# THEORY AND PRACTICE OF Psychological Testing



*Revised Edition*

FRANK S. FREEMAN
*Cornell University*

HOLT, RINEHART AND WINSTON, INC. NEW YORK

Library of Congress Catalog Card Number: 55-6048

22929 0215
Printed in the United States of America

# PREFACE TO THE REVISED EDITION

THIS revised edition, while essentially the same in form as the original one, has been improved and brought up to date where possible The following special features might be noted In this edition there is a fuller discussion of test standardization, particularly as regards metbods of estimating reliability and validity Some of the more significant recently published tests have been included The treatment of projective techniques has been extended in such a manner as to be particularly useful to students who are not specializing in clinical psychology Also, the discussion of tests of specific aptitudes has been extended Throughout, an effort was made to incorporate in discussions and evaluations the results of representative researches that have appeared since the publication of the first edition of this book One other point in particular, should be noted here namely, that considerably more attention and emphasis are given in this edition to psychological analysis of functions being tested by each of the several types of measuring devices This aspect of the subject was not neglected in the first edition, but it has been enlarged in this revision This is not to say that factorial analysis is disregarded, it signifies, however, that the value of such analysis rests basically upon the psychological insights of the test builder at the outset

Throughout this edition, more so than in the first, emphasis has been placed upon the necessity of interpreting test results in the light of the psychological principles involved, of the statistical bases in test construction, and of an understanding of developmental and behavioral principles It is the hope of many of us in this field of psychology that through such emphasis mechanical use and rule-of thumb interpretation of tests will be discouraged, while, on the other hand, the importance of general competence in psychology is stressed

# PREFACE TO THE REVISED EDITION

THIS revised edition, while essentially the same in form as the original one, has been improved and brought up to date where possible The following special features might be noted In this edition there is a fuller discussion of test standardization, particularly as regards meth ods of estimating reliability and validity Some of the more significant recently published tests have been included The treatment of projective techniques has been extended in such a manner as to be particularly useful to students who are not specializing in clinical psychology Also, the discussion of tests of specific aptitudes has been extended Throughout, an effort was made to incorporate in discussions and evaluations the results of representative researches that have appeared since the publication of the first edition of this book One other point in particular, should be noted here namely, that considerably more attention and emphasis are given in this edition to psychological analysis of functions being tested by each of the several types of measuring devices This aspect of the subject was not neglected in the first edition, but it has been enlarged in this revision This is not to say that factorial analysis is disregarded, it signifies, however, that the value of such analysis rests basically upon the psychological insights of the test builder at the outset

Throughout this edition, more so than in the first, emphasis has been placed upon the necessity of interpreting test results in the light of the psychological principles involved, of the statistical bases in test construction, and of an understanding of developmental and behavioral principles It is the hope of many of us in this field of psychology that through such emphasis mechanical use and rule-of thumb interpretation of tests will be discouraged, while, on the other hand, the importance of general competence in psychology is stressed

The major deletions are the chapters on "Statistics in Psychological Testing" and on "Applications and Problems" (Chapters 2 and 16, respectively, in the first edition) These chapters have been omitted to make room for essential additions and elaborations germane to the psychological tests themselves Instructors who found Chapters 2 and 16 of the first edition useful will probably be able to refer their students to that volume for these materials

*January, 1955* *F. S F.*
*Ithaca, New York*

# PREFACE TO THE FIRST EDITION

AN EXAMINATION of the volumes by Guy M Whipple, *Manual of Mental and Physical Tests* * will reveal the changes and developments that have taken place in the field of psychological testing in the last forty years At the same time, such an examination will also reveal the extent of the indebtedness of modern testing practices to the work of Whipple, his contemporaries, and his predecessors When Whipple wrote his *Manual*, he did so in order to bring together, for the first time, a comprehensive and balanced description of psychological tests, representing what he called the "simpler" and the "complex" processes

In the developing scientific field of psychological testing, there is recurrent need for periodic presentation of comprehensive descriptions of tests This volume is intended to meet such a need At the same time, however, I have not limited this volume only to descriptions of these psychological instruments

Shortly after having begun, some years ago, to teach courses in psychological testing, individual differences, and clinical procedures, I was convinced that clinicians and other users and interpreters of test results must have an understanding of the theoretical principles and assumptions upon which tests are constructed For that reason I have at several points in this volume presented basic theories and principles, independent of any specific test or group of tests, theories and principles which are common to a wide range of tests In addition, other theoretical principles, assumptions, and problems have been presented, where relevant, in conjunction with the descriptions and evaluations of particular tests or with several belonging to the

* Originally published in 1910 Revised and published in two volumes in 1914 and 1915 (Baltimore, Warwick and York)

same category of instruments. The mere discussion of basic theories, entirely separate and distinct from descriptions and evaluations of specific testing devices, is a relatively barren procedure for anyone except those who are already experienced and qualified in the subject. It is for this reason that I have described many representative tests in detail, both as to psychological and statistical aspects. The student will thus have specific substance to project against and compare with theories previously presented. Knowledge of how to interpret tests is best achieved through combined understanding of theory and familiarity with test content.

Probably no two authors would be in complete agreement as to tests to be included, though their selections would have many in common. I believe, however, that the devices included in this book are representative of the sounder instruments currently available, though in a few instances poor tests have been mentioned or described for the purpose of illustrating a relevant aspect of the subject. The tests have been so grouped, it is hoped, as to prove most useful to those interested in particular types or levels.

Some historical and developmental background of the subject is provided, especially the work of Alfred Binet, for I believe the student achieves a fuller and sounder appreciation of the present status of the science and its applications through a presentation of early work and thinking. For the advanced student, however, the historical and developmental background provided herein should not suffice; he should consult more detailed and comprehensive studies.*

This volume is intended primarily for students who plan to enter professions in which psychological tests are administered and the results interpreted in dealing with adjustment and numerous other psychological problems. Thus it is designed for the use of clinical psychologists, school psychologists (who are also clinicians), guidance counselors, teachers, psychiatrists, pediatricians, social workers, and personnel officers. Without at all minimizing the usefulness of group testing and group studies of psychoeducational, psychosocial and purely psychological problems, the emphasis herein is on individual and clinical interpretation of test findings. While recognizing that an individual does not exist in a social vacuum and that his

* E.g., Jos. Peterson, *Early Conceptions and Tests of Intelligence*, 1925 (Yonkers, N. Y., World Book Co.), E. J. Varon, "The Development of Alfred Binet's Psychology," *Psychological Monographs*, No. 207, 1935.

behavior and performances can be fully understood only in terms of himself and of his environments, the fact is, nevertheless, that in psychological testing the ultimate unit of concern to the examiner usually is the individual subject After the results of testing are available, in an individual instance, the psychologist may then seek causal and explanatory factors

While formal preparation in statistics, especially in statistical reasoning, is desirable, even in the case of students beginning the study of psychological testing, it is not possible to provide that preparation in a book such as this But understanding of the more common statistical indexes, methods, and reasoning is necessary Hence, a chapter on "Statistics in Psychological Testing" has been included for those students who have not had formal preparation I am pleased to acknowledge my indebtedness to my colleague, Professor T A Ryan, who collaborated by assuming major responsibility for this chapter which is within the area of his major teaching interests

Psychologists will note, of course, that a large and important segment of psychometric methods and devices has been almost entirely omitted from this book, that is, the psychophysical methods of Weber Fechner, Muller, Urban, and their successors The reason for the omission is that psychophysical measurements comprise an area in themselves and are beyond the scope and purpose of this volume

Any author of a textbook is indebted, of course, to the many scientists and scholars who have preceded him and contributed the materials from which the book is developed In specific instances of indebtedness, I have acknowledged authors and sources at each point, where the documentation is most useful In particular, I wish to thank the authors and their publishers who kindly gave permission to reproduce textual, tabular, and graphic materials

One or more chapters were read in manuscript by Dr Solomon Machover, Professor Max L Hutt, and Professor Frederick L Marcuse I wish to acknowledge my appreciation of their valuable criticisms and suggestions

*F S F.*

*Cornell University*
*November 17, 1949*

# CONTENTS


CHAPTER PAGE

1. BASIC THEORETICAL PRINCIPLES 1

*Objectivity in Administering and Scoring A Representative Population Sample Sampling of Traits and Functions Steps in the Development of a Test Reliability Validity*

2. INTERPRETATION OF TEST SCORES QUANTITATIVE AND QUALITATIVE 42

*An Index of Relative Rank Psychological Measurement Contrasted with Physical Measurement Clinical Aspects Difference between Norms and Standards Factors in Selecting a Test*

3. DEFINITIONS AND ANALYSES OF INTELLIGENCE 60

*Definitions of Intelligence Two Comprehensive Definitions Implications for Test Design and Content Three 'Kinds' of Intelligence Analyses of Mental Ability Factor Analysis Illustrations of Factors Implications*

4. THE BINET SCALES 96

*Historical Background The Early Work of Alfred Binet The 1905 Binet-Simon Scale The 1908 Binet-Simon Scale The 1911 Revision of the Binet Scale Summary*

CHAPTER PAGE

5. EARLY REVISIONS OF THE BINET-SIMON SCALE 114
*Four Early Revisions The Stanford Revision of 1916*

6. THE 1937 REVISION OF THE STANFORD-BINET SCALE 129
*Description of the 1937 Scale Validation Reliability Determining Mental Age and Intelligence Quotient Distribution of IQ s Suggested Classification of Revised Stanford Binet IQ s Analysis of Functions Tested Types of Items The Short Scale Evaluations and Criticisms*

7. THE WECHSLER SCALES 156
*Description of the Wechsler Bellevue Intelligence Test Functions Involved in the Subtests Need for an Adult Scale Standardization The Population Sample Validity Reliability Scoring and IQ Calculation Special Features of the Bellevue Scale Criticisms and Evaluations The 1955 Revision of the Bellevue Scale The Wechsler Intelligence Scale for Children (1949)*

8. INDIVIDUAL PERFORMANCE SCALES 197
*The Pintner Paterson Scale of Performance Tests The Cornell Coxe Performance Ability Scale The Arthur Point Scale of Performance Tests Revised Arthur Scale Form II Other Performance Tests Functions Tested by Performance Scales Evaluation of Performance Tests*

9 SCALES FOR INFANTS AND PRESCHOOL CHILDREN 220
*Gesell Developmental Schedules Minnesota Preschool Scale Cattell Developmental and Intelligence Scale Merrill Palmer Scale of Mental Tests Evaluation of Scales for Infants and Preschool Children*

10. NONVERBAL GROUP SCALES OF MENTAL ABILITY 241
*Beginnings Characteristics of Group Tests of Mental*

CHAPTER PAGE

*Ability Pintner Cunningham Primary Test Chicago Nonverbal Examination Revised Army Beta Examination Pintner Nonlanguage Series Intermediate Test Nonlanguage Multi Mental Test Pattern Perception Test Progressive Matrices Test Cattell Culture–Free Test Goodenough Drawing Test Davis Eells Test of General Intelligence Evaluation of Nonverbal Group Scales*

11 VERBAL AND MIXED GROUP SCALES OF MENTAL ABILITY 270

*California Tests of Mental Maturity Terman McNemar Test of Mental Ability Tests of Primary Mental Abilities Kuhlmann Anderson Tests (6th Edition) Group Scales for College Freshmen Army General Classification Test Miller Analogies Test Other Group Scales Evaluation of Group Scales Uses of Group Scales*

12 APTITUDE TESTS MECHANICAL AND CLERICAL 306

*Definition and Explanation Tests of Vision and Hearing Motor and Manual Tests Tests of Mechanical Aptitude Tests of Clerical Aptitude Differential Aptitude Tests Aptitude Classification Tests*

13 APTITUDE TESTS FINE ARTS AND PROFESSIONS 336

*Tests of Musical Aptitude Tests of Aptitude in the Graphic Arts Tests of Aptitude in Medicine Tests of Aptitude in Law Tests of Aptitude for Teaching Tests of Science and Engineering Aptitudes Interest Inventories General Evaluation of Aptitude Tests Steps in an Aptitude Testing Program*

14 TESTS OF EDUCATIONAL ACHIEVEMENT 377

*Scope Purposes Derived Indexes Types of Items Three Representative Batteries Reading Tests Arithmetic Tests Tests at High School and College Levels Tests of Aptitude in Specific Academic Subjects Tests of More Complex Educational Objectives Evaluation of Achievement Tests Tests of Proficiency*

CHAPTER PAGE

15 INTELLIGENCE TESTS AS CLINICAL INSTRUMENTS 403

*Factors Which Affect Test Performance The Stanford Binet Scale The Bellevue Scale Kent Series of Emergency Scales A Report Outline*

16 TESTS OF MENTAL IMPAIRMENT 432

*The Babcock Test Tests of Concept Formation The Hunt Minnesota Test for Organic Brain Damage The Bender Visual Motor Gestalt Test Evaluation of Tests of Impa rment*

17 PERSONALITY RATING SCALES 452

*Defin tion of Personality Rating Scales Major Aspects Representative Rating Scales Evaluation of Rating Scales*

18 PERSONALITY INVENTORIES 466

*Purposes and Types of Inventories Representative Inventories Biographical Data Questionna res Tests of Attitudes and Values Opinion Polling Evaluation of Personality Inventories*

19 PROJECTIVE METHODS THE RORSCHACH AND THE THEMATIC APPERCEPTION TESTS 502

*Defin tion and Explanation The Rorschach Test Thematic Apperception Test*

20 PROJECTIVE METHODS VARIOUS 547

*Word Association Tests Pictu re Tests Verbal Com pletion Tests Drawing and Painting Play Evaluation of Projecti e Tests*

21 SITUATIONAL TESTS 577

*Sociometric Methods Tests of Social Intelligence and Leaders ip Psychodrama Office of Strategic Services Assessment Tests Evaluation of Situational Tests*

INDEX 599

# I.

## BASIC THEORETICAL PRINCIPLES

ALTHOUGH tests of general intelligence, specific aptitudes, personality, and educational achievement are designed and constructed for different purposes, all of them have certain principles and procedures in common, and any combination of these categories of tests might be used in dealing with a specific individual case or in attempting to solve a particular psychological problem Psychological tests have been used to find answers to a number of psychological questions both theoretical and practical But ultimately, and most important, they are intended to contribute to the analysis and description of individuals, and to the evaluation, prediction, and guidance of their behavior and education The following are the aspects that are common to the several types of psychological instruments and that give them their objectivity objectivity in administering and scoring, norms based upon a population sampling, scientifically selected, for a particular test, sampling of specified traits or functions, by means of a particular test, incorporation, within a test, of a composite of views of a number of experts, utilization of recognized techniques of test standardization The tests' objectivity and the standardization process give them a scientific quality which, of course, is absent in an individual's personal estimate of psychological traits and functions *We define a psychological test as a standardized instrument designed to measure objectively one or more aspects of a total personality, by means of samples of performance or behavior*

## OBJECTIVITY IN ADMINISTERING AND SCORING

Each psychological test is administered under a prescribed set of procedures These procedures involve preparation of the persons to be tested by means of introductory and explanatory remarks, phrasing and formulation of instructions whereby each part, or each item in some instances, is to be presented, setting time limits, if any, decisions as to when to repeat instructions or to offer encouragement, and when not to do so, as well as when to answer questions asked by a subject, and when not, the use of practice exercises, if any

The scores and ratings thus derived from the tests are not dependent upon the individual bias or judgment of the particular examiner For the score of any subject on an objective test is arrived at by the use of a scoring key, or the scoring is otherwise so clearly defined, specified, and illustrated that subjective judgments of individual examiners or scorers do not enter in at all or are reduced to a minimum Thus an objective test provides a highly uniform means of evaluating the psychological traits or functions being measured, results obtained by one competent examiner are comparable with those obtained by others

## A REPRESENTATIVE POPULATION SAMPLE

Every test is designed and intended for use with a specified *population*, or group For example, a test of intelligence may be standardized for use with individuals from the age of two years through adulthood (Stanford Binet, 1937 revision), another for ages eleven to seventeen (Chicago test of Primary Mental Abilities), another primarily for adults (Wechsler-Bellevue) Still others cover different age ranges

A test of scholastic achievement in a particular school subject or group of subjects may be intended for the first three grades, or for grades eight through twelve, or for college freshmen, or for other grade ranges depending upon the school subject and the prescribed scope

Tests of specific aptitudes likewise are designed for specified populations For example, one test of ability in art is designed for grades seven and above, one test of mechanical aptitude is to be used for ages eight to twenty one, a law aptitude test is standardized, of course, for

college students and others who are candidates for admission to a law school, regardless of age

Rating scales, personality inventories, and projective tests are likewise intended for use with a specified segment of the total population They may be designed for a selected age group, for particular occupations, for given educational levels, for one sex of limited age range, for the diagnosis of clinical cases, or for use with non-clinical populations as well

In any event, whatever the traits or functions to be measured, whatever the range of ages or school grades, and whether for clinical or non-clinical groups, the test must be standardized upon a group that is a representative sample of the total population for which it is intended Each test must be constructed by means of actually sampling the performance of an adequate group which has been selected in such a way as to insure its being typical of the population of which it is a part

Factors to be taken into account in making a population sampling will depend upon the nature and the comprehensiveness of the test under construction In any instance, the sample should yield unbiased data on the population of which it purports to be representative, and the sample should be large enough to provide statistically valid results for the traits or functions being measured by the test

This means, of course, that the author of a test must decide at the outset with which group, with what segment of the population his instrument is to be used Then he must standardize his test upon a population sample that is stratified according to relevant factors, and within each stratum the selection of cases should be adequate in number and of correct proportion in the total [1]

For example, if a psychologist is to construct a test of 'general intelligence' for American children, in the primary grades ranging in age from five to nine years, he will have to take into account the following factors in obtaining his standardization population age, sex, geographic area, parental occupational level and type of community (urban, village, farm) The author of the test must decide, also, whether he will standardize his test entirely on a Caucasian population, or whether he will include non Caucasian elements If it is to be

[1] On sampling see M B Parten *Surveys Polls and Samples* New York Harper, 1950

the latter, then the racial factor must be taken into account in the stratification of the sample

Since individuals within a representative sample of children of any given age vary widely in respect to mental abilities, some will reach only the levels of younger age groups while others will attain the levels of older age groups Thus, to ascertain the developmental level of the retarded it is necessary to extend downward the chronological age of the standardization sample, and, conversely, it is necessary to extend upward the age limit for the superior

The validity of results obtained with any psychological test will be dependent, in part, upon the adequacy and representativeness of the standardization population

## SAMPLING OF TRAITS AND FUNCTIONS

Any given test measures a limited aspect of the person being examined, though some tests are much more restricted in scope than others It is essential, therefore, that the test builder define the aspect, or aspects, he proposes to measure After doing this, he must develop a series of test items that will best sample the traits or functions with which his test is concerned

In developing a psychological test, it is impossible, and in fact unnecessary, to use an unlimited number of items It is not necessary to attempt to present the individual being tested (called the "subject" or the "testee") with problems that will ascertain his responses for every conceivable situation involving a given trait or function It is sufficient to get an adequate sampling of responses in a particular area or range of behavior, the assumption being that the sampling is representative of the whole

Two kinds of sampling are actually involved in constructing a psychological test First, the most relevant constituents of the *gross variable* (the broad, comprehensive trait or function) must be selected Where, for example, the gross variable is "general intelligence," the constituent parts in the test might be vocabulary, verbal comprehension, arithmetical problems, reasoning with practical problems, verbal and other analogies, perceptual organization, and so forth Second, the *operational levels* (that is, the actual items) must be selected which arithmetical processes and at what levels, what kinds and which levels of words, what types and range of situations, which perceptual figures?

In following this procedure, psychologists are employing a well-

known and widespread technique If a chemist wishes to determine the quality of a shipment of milk, he takes small quantities here and there, combines these, and then analyzes a sample of the samples If an agronomist wishes to analyze a given area of soil he gathers small amounts from various spots If a blood test is to be made, a very small quantity, taken from one place, is sufficient and representative of the entire stream Numerous other illustrations can be found So, too, with intelligence, specific aptitudes, and school achievement It has been said that psychological testing may be thought of, figuratively, as sinking shafts here and there within a given range in order to ascertain depth and quality

Specifically, for present purposes of illustration, *intelligence* may be defined in several ways (1) capacity to integrate experiences and to meet a new situation by means of appropriate and adaptive responses, (2) capacity to learn, (3) capacity to carry on abstract thinking [2] While psychologists differ in regard to which of these three aspects is most important and which they would emphasize, the fact is that most tests of general intelligence probe and sample all three The following types of items found in various current tests fall under one or more of these definitions, and are constituent parts of the gross variable, general intelligence

Practical reasoning What s the thing for you to do when you have broken something which belongs to someone else? (From the Revised Stanford Binet Scale, Form L )

Definitions of words i e, concept formation

Perceiving similarities and differences between objects For example In what way are wood and coal alike? In what way are a baseball and an orange alike and in what way are they different? (From Revised Stanford Binet Scale, Form L ) i e, abstraction and generalization

General information tests i e assimilation and retention of experiences

[2] Intelligence will be defined and described in a later chapter There are some psychologists who prefer to discard the term *intelligence* because they believe it is not a function in itself but rather an aggregate of particular aptitudes or, to use a more recent term of primary mental abilities Like many other psychologists we continue to use the term *intelligence* which we prefer for two reasons (1) It has general and meaningful currency now, especially in connection with tests (2) Even if certain abilities were primary" intelligence would not be a mere aggregation of these but rather an integration in which case intelligence is really something new different from and more than its several constituent parts

Arithmetical reasoning i e, reasoning with abstractions
Supplying missing parts to pictures i e, perceptual integration
Reproducing geometric figures from memory i e, visual imagery and organization
Arranging a series of pictures in logical sequence i e, visual perception and reasoning
Perception of color and form design i e visual imagery and recall, analysis and organization
Explanation of absurdities in given pictures i e, analysis of visual percepts
Oral solution of practical problems orally presented i e, analysis and generalization
Solving problems involving distances and directions (without use of paper and pencil) i e, spatial orientation
Deriving and giving the meanings from a prose passage i e, reasoning with abstractions

Another method of determining the component parts of the gross variable, which in this case we call *general intelligence*, is through "factor analysis," to be discussed more fully in Chapter 3 According to one analysis, there are six such components, relatively independent of one another [3] facility with numbers (the four fundamental processes), vocabulary (word meaning), space perception (perceiving similarities of and differences between geometric figures), word fluency (controlled word association), reasoning (insight into patterns of letters arranged in series), and memory (immediate recall of discrete verbal materials) Tests have been constructed on the basis of this analysis, items having been devised for each of the six categories

At present, also, there is a trend among some psychologists toward analysis of the gross variable's component parts into subdivisions, that is, into *elements* of the component parts For example, the component "reasoning" has been tentatively analyzed into the following four aspects [4]

Reasoning I
- a manipulating symbols
- b solving problems
- c *defining problems*
- d testing hypotheses

[3] L L Thurstone and T G Thurstone *The Chicago Test of Primary Mental Abilities* Chicago Social Science Research Associates, 1943

[4] J P Guilford et al *A Factor Analytic Study of Reasoning Abilities* Los Angeles University of Southern California Report Number 1, June 1950

Reasoning II
- a seeing rules or principles (induction)
- b seeing systems
- c seeing trends
- d seeing relations (educing relations)
- e seeing identity of relationships
- f analyzing forms

Reasoning III
- a seeing common elements or properties
- b classifying (in general)
- c classifying forms
- d educing correlates

Reasoning IV
- a drawing inferences (deduction)
- b syllogistic reasoning

Inspection of these four types of reasoning reveals that they are neither mutually exclusive nor independent of one another Yet if these and their parts are sufficiently distinct and constitute reasoning in its several aspects, and if reasoning were to be measured according to this scheme, it would be necessary to devise items for each of the four types and for each sub type

*Specific aptitude* as another example, may be defined as a capacity that indicates the probable degree of successful learning and achievement in a particular and limited type of activity—for example, musical mechanical artistic, or linguistic aptitude A test intended to estimate a person's capacity in each of these must include parts and items sufficient in number and extensive enough in scope to provide an adequate sampling upon which a prediction of subsequent learning and achievement may be based

The constituent parts of a test of 'mechanical aptitude,' for instance, might be knowledge of tools and mechanical devices, skill in assembling parts of a mechanism, perception of spatial relations, manual or digital dexterity, or others found through statistical and psychological analysis to have predictive and selective value

*Personality* tests and inventories must also be based upon samplings of the constituent traits that the test author proposes to evaluate This is true even though personality itself is most difficult to define and though its components are elusive Thus we have inventories that attempt to measure, among others degrees of introversion-extroversion, neurotic tendencies, anxiety, hypochondriasis, adjustment to home, adjustment to school, and dominance submission Each author of a

nality inventory or of a projective test, in order most adequately to fulfill his purposes, must determine which aspects of personality are to be examined by means of his instrument

An *educational achievement* test is designed to measure an individual's information in, or skill with, or understanding of—or all three of—a given subject of study taught in school, for example, reading rate and comprehension, arithmetical processes and problem solving, American history, English usage, and so on In each instance, as in all other types, the scope of the test must be defined, the parts of the gross variable must be determined, and the elements of each part must be represented Educational achievement tests depend for their validity upon the adequacy with which they sample the subject matter field for which they are intended

Tests of general intelligence, of specific aptitudes, and of educational achievement are intended, of course, to indicate the person's status, at the time of examination, in their respective areas, but they are intended, also, for other purposes Intelligence tests are employed to predict an individual s probable future level of mental development and capacity Tests of specific aptitude are used to predict probable future learning of and performance in a particular activity or occupation Results of educational achievement tests are helpful in forecasting the subject s probable future level and quality of learning in the several school areas and in diagnosing specific difficulties and disabilities in basic school subjects The great importance, therefore, of adequately conceived and satisfactorily standardized instruments is readily apparent

In respect to determination of an individual s present status, personality inventories and projective tests are like the three other types mentioned above And while personality tests are used, to some extent, for predicting future behavior and adjustment, their greater significance and usefulness, in relation to one's future status, lies in the fact that their results are very valuable in diagnosing personality problems and as a basis for psychological counseling, or therapy where indicated.

## STEPS IN THE DEVELOPMENT OF A TEST

In devising a test it is necessary, as already explained, first to define that which is to be measured Tests of intelligence, specific aptitude, school achievement, and personality have already been briefly

defined, and illustrative materials will be presented in later chapters At this point, several illustrations will suffice to demonstrate how psychological and statistical analysis determine the form and content of a test

Alfred Binet, the French psychologist of whose work much more will be said later, proposed that the following be tested as components of intelligence memory, mental images, imagination, attention, comprehension, suggestibility, esthetic appreciation, muscular strength, strength of will, motor skill, and visual judgment Some of these survived his own experimental investigations and those of other psychologists, while others were rejected as being invalid Binet suggested the measurement of these processes in the first place because he believed that they differ sufficiently from person to person and that knowledge of their level of development in different persons would permit the psychologist to distinguish one person's general mental development from that of others Here we have the beginnings of the tests which were to prove so useful in the construction of his and other scales

Subsequently, psychologists were to identify other and more specific processes which, they held, should be tested Thus Spearman developed the theory that intelligence is essentially a generalized function (*g*) and should be measured through a broad sampling of mental activity E L Thorndike, on the other hand, maintained that intelligence consists of a multitude of highly specific processes (unnamed) and that the validity of a test of intelligence will depend upon sampling of these with psychological insight L L Thurstone, among others, developed still a different theory namely, that what we call "intelligence" is made up of a number of "primary mental abilities," each of which must be sampled, with a "secondary" general factor involved These theories and other problems concerning the nature of intelligence will be dealt with in Chapter 3, for the present we are concerned only with definitions and analysis of processes as necessary steps in the development of tests

The Seashore tests in music include the following processes pitch discrimination, judgment of tone intensity, perception of time, discrimination of tonal timbre, tonal memory, and rhythm discrimination

In testing knowledge of and skill in language, any or all of the following might be included grammar, usage, punctuation, vocabulary, reading rate, reading comprehension, visual acuity, eye movements, and auditory acuity

Regardless of the particular definition of intelligence that commends itself most to a psychologist and regardless of what analysis into particular processes appears to have greatest value, any psychological test must measure the trait and function in its *manifestations* in one form or another The task then is to devise and select items conforming to the definition of the function or trait and to the analysis made of it

After the original group of items has been devised and selected, they are given a series of try-outs on the groups for which they are intended The results thus obtained are subjected to established statistical analyses and to scrutiny on the basis of a series of criteria of validity and reliability As a result of this first analysis and scrutiny, some items are rejected, some are retained, and new ones are added The new and more highly selected items are again put through the same process of *validation* resulting in further improvements and refinements of the test under construction This will be repeated several times before the finished scale emerges The entire process of try-outs, statistical analysis and evaluation on the basis of accepted criteria is called *test standardization*. Standardization of tests is a topic which will recur with some frequency in this volume For the present, however, it will suffice to indicate what are the most important aspects involved in the process, in addition to those already discussed

**RELIABILITY**

The two essential characteristics of any sound test are *reliability* and *validity*

The term *reliability* has two closely related but somewhat different connotations in psychological testing. First, it refers to the extent to which a test is *internally consistent* that is, the extent to which the test scores are subject to or free from such internal defects as will produce errors of measurement due to the quality of the items rather than to the instability of performances of testees themselves In other words, how accurately is the test measuring at a particular time? Second, *reliability* refers to the extent to which an instrument yields consistent results on testing and retesting That is, how dependable is it for predictive purposes? Obviously, if a test does not have a high degree of reliability, it can have but limited value, if any, in predicting an individual s future performance or level of development It is clear that these two aspects of reliability are intimately related if a test is

not highly reliable on any particular occasion, it can have little predictive value

Since one of the principal uses of psychological tests is to predict and plan for subsequent development and performance, a high degree of reliability is a *sine qua non* of a sound instrument Reliability is not, however, an all or-none proposition, it is a matter of degree No test is perfectly reliable, the scores obtained on repeated testings are not completely stable, either in terms of internal consistency or of prediction There are always some errors of measurement, large or small, and it is "normal" for humans to vary in performance, generally within fairly narrow limits, from one occasion to another—to vary, that is, aside from the expected changes that occur as part of the process of growth and development

Differences among the test scores of individuals in a group are due to (1) "true," or actual, differences in the trait being measured within those persons being examined, and (2) sources of inaccuracy in the measurement of individuals These sources of inaccuracy may be inherent defects of the test itself, conditions or "chance" factors operating at the time of testing, or unpredictable fluctuations in the performance of the subjects Standardization aims to eliminate or reduce inherent defects of the test, the conditions of testing and retesting should be as nearly consistent and optimal as possible, and though minor fluctuations in an individual's performance from day to day or week to week cannot be controlled, the reasons for any major fluctuations must be sought in the individual himself or in some of the environmental forces, if the sources of inaccuracy are to be understood

The possible sources of variation in performances on a test are many This aspect of testing will recur frequently in our subsequent discussions, but for the present, the most common of them may be listed as follows

Actual or true," differences among individuals in the *general* traits or *general* abilities being measured

*Specific* abilities required in a particular test, or *specific* disabilities in the functions being tested

Skill in taking tests, being test wise,' or the converse

The 'chance' acquisition of a particular piece of knowledge or information required in a test e g, the meaning of an unusual word

such as *ambergris* or a bit of unusual information such as the name of the author of a little known work (These would be poor test items )

Effects of practice (previous test taking) or, in some instances, coaching

Normal or expected fluctuations in performance from time to time

Personal characteristics of the testee motivation, health energy level, emotional status

Physical conditions under which the test is taken heat, light, ventilation

Unpredictable, or chance' factors noise, interference, broken pencil, misunderstanding of instructions, etc

Fortunate guessing of answers

Test results, ideally, should depend upon the extent to which the test measures the first two of these sources of variation, actually, however, the coefficients of reliability will be adversely affected by the nonsystematic operations of the others

There are two methods, in general, of expressing the consistency, or dependability, of test results (1) *absolute reliability*, and (2) *relative reliability* The first of these is usually stated in terms of the *standard error of measurement* The second is given, though infrequently, in terms of *analysis of variance*, or, much more commonly, as a *correlation coefficient* indicating the degree to which individuals maintain relatively consistent positions in their group when a single test is administered twice, or when two equivalent forms of a test are applied to all members of a group This correlation between the two sets of scores is known as the *coefficient of reliability*

**Test-retest reliability.** When persons are tested and retested a number of times, they may undergo some change as a result of repeated measurements e g, in the form of practice effects, improvement in the skill of taking tests, and in the 'set' or attitude toward a test In estimating reliability, therefore, it is necessary to limit the number of times an individual is tested with the same device Hence, instead of frequent retesting of the same persons, dependable results for a given psychological instrument are obtained by increasing the number of persons tested rather than by increasing the number of measures of

each person Therefore, techniques have been devised for evaluating the results obtained with only one or two measurements of the same individuals, namely

Two equivalent forms of the test are administered and the two sets of scores are correlated

A single form of the test is administered twice and the two sets of scores are correlated

The items of a single test are subdivided into two separately scored groups the two sets of scores being correlated as though they were obtained from two forms or two testings

Administering two equivalent tests has several disadvantages The procedure requires more time, of course The two forms might vary somewhat in content, thus underestimating reliability of either form The experience of having taken the first test might result in some learning or improved skills If two forms of a test are to be equivalent, they must be in the same format and each must test a representative sampling of items measuring the same mental processes Also, the original testing and the retesting should take place within a week or two in order to minimize the influence of intervening factors of developmental and other individual changes

Administering the identical test twice has some of the disadvantages of using two equivalent forms It is held by some investigators that recall of answers to specific items of a test is an added disadvantage when the identical test form is given a second time Although there can be some recall, it is unlikely that this possibility will be an important consideration, for the number of items in any test is too large for the retention of many When this method of estimating reliability is used, the interval between testings should be a week or two in order to minimize the effects of whatever recall might be operative

**Split-half reliability.** A test cannot have high consistency in retesting unless each application is relatively free of chance or random errors In *split half* testing of reliability, chance and random errors may be assumed to operate equally in both halves

Calculating reliability by the ***split half*** method consists of subdividing the whole test into two parts, presumably equivalent, and then treating the score of eacb part as though it were a separate form This method provides, essentially, a measure of the test's internal consistency, assuming an equal level of performance throughout the test by each person Split half reliability is a first check upon the usefulness

of a test It is easily found and saves unnecessary labor that might be spent in following up an internally unsound device This method tells us if the test is a reliable representation of an individual's traits at a given time It does not describe completely the reliability of a test which is to be used periodically or for predictive purposes For periodic and predictive testing, the test-retest method is desirable

The split half method of determining reliability may, in some circumstances, yield a coefficient of correlation that is somewhat too high In calculating reliability, an assumption is that the operations of chance factors are uncorrelated and hence will cancel out one another But in using the split-half method, both obtained measures are determined at the same sitting and any chance fluctuations due to temporary conditions within testees and to conditions in the external situation will operate in the same direction and thus yield a somewhat higher correlation coefficient than might be found by other methods

Generally, for split half reliability, the subdivision is made by taking the odd numbered items as one part of the test and the even-numbered items as the other The score is then found for each person, for each of the subdivisions (This method is referred to as *odd even* reliability ) Since the correlation coefficient for the two sets of scores derived by this method is based upon subdivisions of the full test, each of which is half the length of the whole, a statistical formula (Spearman Brown) is used to correct for the reduced lengths of the subdivisions from which the correlated scores have been determined The reason for this correction is that the score of the whole test, being based upon a larger number of items, is a more adequate sampling of traits or functions and hence reduces the possible effects of chance solutions and accidental errors The whole test is thus more reliable than its subdivisions, and the correction formula is intended to indicate what the reliability of the entire test would be, based upon what was found with the part scores

An example will demonstrate how the Spearman Brown formula operates The generalized formula is

$$r_n = \frac{nr}{1 + (n-1)r}$$

in which $r$ is the coefficient of reliability obtained between the parts of the divided test, $r_n$ is the reliability of the test n times as long as half the original test

In the method of odd-even reliability, $n$ is 2, since the original test

has been divided in two equal parts Assuming, then that the odd-even coefficient ($r$) is 80, and substituting the values in the formula, the reliability of the whole test ($r_n$) is found to be 89 This estimated reliability coefficient for the test as a whole is the one usually reported in psychological research and in test manuals

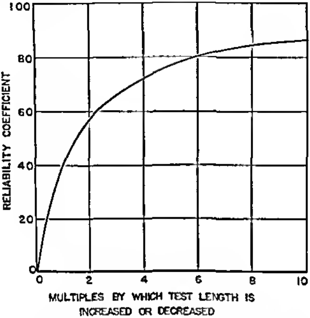



FIG 1 1 *Changes in Reliability with Changes in Length of Test, as Predicted by Spearman Brown Formula Unit Length Reliability is 40* From L L Thurstone, *The Reliability and Validity of Tests* Ann Arbor, Mich, Edwards Bros, 1935

The Spearman-Brown formula may be used to estimate the effect upon reliability of a test of a given length if it should be increased by any multiple (say, 3 or 4 times) or decreased by any fraction (say, ½ or ⅓) There is a point of diminishing returns, so to speak, beyond which the very small increase in the reliability coefficient, resulting from increase in length, does not warrant the extension of a test (See Figure 1 1) Figure 1 2 illustrates increase in test reliability as the length of a test is doubled This figure demonstrates what happens when reliability is calculated by the split half method and then corrected by the Spearman-Brown formula

Selecting odd-numbered items as one half of the test and even-numbered items as the other half is justified on these grounds: items in most tests (as will be seen) are grouped together according to type (number sequences, vocabulary, etc.) and are graduated according to difficulty, from easiest to hardest. Thus, when this systematic arrangement is employed, the odd-even procedure yields very close approxi-

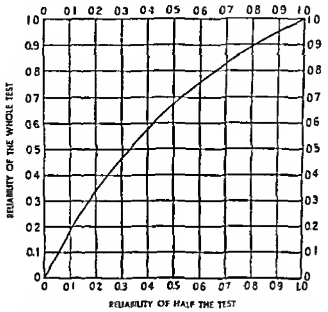



FIG 1 2. *Showing the Increase in Reliability of the Whole-Test Scores as a Function of the Reliability of Half-Test Scores, when the Spearman Brown Formula Is Applied.*

mations to equivalent half-scores, because each half-score is based upon the same types of items and the same number of each type; and each half-score is based upon items which progress in difficulty in approximately the same degree. For example, consider the first ten items of a single type (known as a subtest), say, verbal analogies Numbers 1, 3, 5, 7, 9 are, as a group, of approximately the same total difficulty as numbers 2, 4, 6, 8, 10—if they are graduated in difficulty from 1 to 10, for both the odd-numbered and the even-numbered include items from practically the entire range of difficulty represented from numbers 1 to 10.

There are other methods of selecting items for getting part scores, but since they are not too frequently encountered, they will not be described here Whatever method is used to get equivalent half-scores, the procedure must be based upon the psychological rationale and upon the format of the particular test under consideration, in order to insure, as far as possible, equivalence of items in respect to mental processes involved and in respect to difficulty, as well as number of items in each part

The split-half method tends to overestimate the *predictive* reliability of an instrument, since the correlation is not affected by the ordinary conditions that cause normal fluctuations in a person's performance on different days In particular, this method should not be used in estimating the reliability of a pure "speed test,"—by which we mean a test whose items are of the same degree of difficulty throughout and which therefore measures only rate of performance at the given level of difficulty Since all items in the test are of equal difficulty, an examinee should do as well with any one item as with any other Hence, to measure rates of performance and to differentiate among individuals, the time limit and the length of the test should be such that no one is able to complete all the items Under the circumstances, except for chance errors in performance, the odd-even correlation should be +1 00 (perfect positive), because the test is, presumably, uniform throughout and the psychological function being measured (speed) is operating uniformly on all items It is apparent, thus that the total scores on the odd numbered items should equal those on the even numbered One test manual, for example, reports an odd-even reliability coefficient of 99+, but the manual also reports a coefficient of 88 when the scores of two equivalent forms were used[5]

The best practice is to use the test retest method with a highly speeded test Tests differ in respect to the significance of speed of performance, even when the items are also scaled in difficulty As the role of speed in a test decreases, the odd-even correlations will differ less and less from those obtained with the test retest method

**The standard error of measurement.** This index is an estimate of the deviation of a set of obtained scores from their "true" scores[6] It is

[5] *Differential Aptitude Tests* Manual page C-6 New York The Psychological Corporation

[6] A "true" score is the measure that is quite free from and uncontaminated by chance factors and errors of measurement theoretically, it represents an individual s true level of performance on the test being used

Selecting odd-numbered items as one half of the test and even-numbered items as the other half is justified on these grounds items in most tests (as will be seen) are grouped together according to type (number sequences, vocabulary, etc ) and are graduated according to difficulty, from easiest to hardest Thus, when this systematic arrangement is employed, the odd-even procedure yields very close approxi-

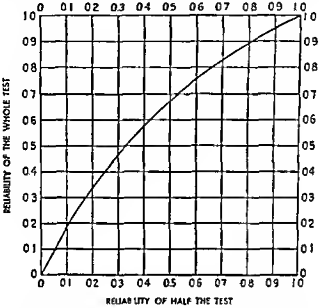



FIG 1 2 *Showing the Increase in Reliability of the Whole Test Scores as a Function of the Reliability of Half Test Scores when the Spearman Brown Formula Is Applied*

mations to equivalent half scores, because each half-score is based upon the same types of items and the same number of each type, and each half score is based upon items which progress in difficulty in approximately the same degree For example, consider the first ten items of a single type (known as a subtest), say, verbal analogies Numbers 1, 3, 5, 7, 9 are, as a group, of approximately the same total difficulty as numbers 2, 4, 6, 8, 10—if they are graduated in difficulty from 1 to 10, for both the odd numbered and the even numbered include items from practically the entire range of difficulty represented from numbers 1 to 10

There are other methods of selecting items for getting part scores, but since they are not too frequently encountered, they will not be described here Whatever method is used to get equivalent half-scores, the procedure must be based upon the psychological rationale and upon the format of the particular test under consideration, in order to insure, as far as possible, equivalence of items in respect to mental processes involved and in respect to difficulty, as well as number of items in each part

The split half method tends to overestimate the *predictive* reliability of an instrument, since the correlation is not affected by the ordinary conditions that cause normal fluctuations in a person's performance on different days In particular, this method should not be used in estimating the reliability of a pure "speed test,"—by which we mean a test whose items are of the same degree of difficulty throughout and which therefore measures only rate of performance at the given level of difficulty Since all items in the test are of equal difficulty, an examinee should do as well with any one item as with any other Hence, to measure rates of performance and to differentiate among individuals, the time limit and the length of the test should be such that no one is able to complete all the items Under the circumstances, except for chance errors in performance, the odd-even correlation should be +1 00 (perfect positive), because the test is, presumably, uniform throughout and the psychological function being measured (speed) is operating uniformly on all items It is apparent, thus, that the total scores on the odd numbered items should equal those on the even numbered One test manual, for example, reports an odd-even reliability coefficient of 99+, but the manual also reports a coefficient of 88 when the scores of two equivalent forms were used [5]

The best practice is to use the test retest method with a highly speeded test Tests differ in respect to the significance of speed of performance, even when the items are also scaled in difficulty As the role of speed in a test decreases, the odd-even correlations will differ less and less from those obtained with the test retest method

**The standard error of measurement.** This index is an estimate of the deviation of a set of obtained scores from their "true" scores [6] It is

[5] *Differential Aptitude Tests* Manual page C-6 New York The Psychological Corporation

[6] A "true" score is the measure that is quite free from and uncontaminated by chance factors and errors of measurement, theoretically, it represents an individual's true level of performance on the test being used

dependent upon the standard deviation of the distribution of obtained scores and upon the coefficient of reliability of the test from which the distribution of scores was obtained The formula for determining the standard error of measurement is written

$$SE_{(meas)} = SD_x\sqrt{1 - r_{xx}}$$

in which $SD_x$ is the standard deviation of the distribution of the obtained scores, and $r_{xx}$ is the reliability coefficient of the test

Assume that the standard deviation of a test ($SD_x$) is 12 IQ points and that its coefficient of reliability ($r_{xx}$) is 90 Substituting these values in the formula, we find that the standard error of measurement is approximately ±3 8 points This statistic is interpreted as follows assuming that the test scores are normally distributed and that the "errors of measurement" are similarly distributed, then approximately 68 percent of the obtained scores are *within* ±3 8 points of the true scores for the persons measured Otherwise stated, the odds are 68 out of 100 (or 68 to 32) that a particular individual's obtained score is in error by 3 8 points *or less* Then using the table of probabilities for standard deviation values, we can say, further, that the probabilities are 19 to 1 (95 in 100) that the error of measurement will be 7 6 points (twice the standard error of measurement) or less, and 99 to 1 that it will be 9 5 points (2½ times the standard error of measurement) or less

The foregoing technique gives us the means of estimating an individual's "true" score from a set of obtained scores, of which his is one Using the data of the illustration above, assume an individual's obtained IQ is found to be 100 The probabilities are, then, two to one that his "true" IQ lies between 96 2 and 103 8 (For practical purposes we would say between 96 and 104 ) And the probabilities are nineteen to one that it lies between 92 4 and 107 6

Obviously, the higher the test's reliability coefficient, the smaller will be the error of measurement, and therefore the greater the predictive value of the test The standard error of measurement provides us, also, with a basis for judging whether or not the scores for two persons represent a true difference or whether they are only deviations from the same, or nearly the same, true scores For example, if one person gets an IQ of 100 and another person gets one of 96, are these within the range of the same true score or are they significantly different, statistically? Using the data of the illustration given above, we say they are within the range, and they are not sta-

tistically significant Also quite aside from any question of statistical significance, a clinical psychologist knows from experience with details of test performance that no *psychological* significance attaches to a difference between IQ s of 100 and 96, or to similar differences elsewhere on the scale

It is clear that while it is essential to have the reliability coefficient for a test, as an estimate of relative reliability it is equally essential to have the error of measurement as an estimate of absolute reliability [7]

**Analysis of variance.** As already stated, the degree of reliability of a test depends upon the extent to which variations in scores of the testees are attributable to 'true' differences among the individuals constituting the group, and the extent of inaccuracies of measurement A test is unreliable in proportion to the variation of results attributable to factors of test inaccuracy, rather than to 'true" differences among the members of the group The estimate, in the scores of a group, of the proportions of variation due to each of the several factors, is technically known as *analysis of variance* [8]

In a study of intelligence test reliability by this method, we would ask what factors may be important, and to what extent, in producing the obtained differences of scores on two applications of the identical test (or of equivalent forms) to the same group of persons? First since individuals differ in any population sample, the analysis should estimate the extent to which obtained differences in scores are due to 'true differences in the functions being measured Second if there is some general improvement of scores on the second test, it would be necessary to estimate the practice effect Third since the two foregoing factors would, in all probability, not account for all differences in scores it is assumed that there are residual differences due to errors of measurement attributable directly to the test being used,

---

[7] The SE (meas ) is an over all index theoretically applicable throughout the range of scores It sometimes happens however that a test measures with less error at some parts of the scale than at others In that case it is possible to determine at which parts of the scale the "errors of measurement" are larger or smaller On the Stanford Binet Scale (1937) for example the SE (meas ) is 5 2 points for IQ s above 130 but only 2 2 points for IQ s below 70

[8] *Variance* is defined as the mean of the squared deviations from the mean score of the group A measure of deviation is an index of the extent to which individual scores of a group vary from the group s average score Variance is the statistical term for the square of the standard deviation

that is, weaknesses or defects within the test If additional influencing factors could be isolated, their significance in producing the obtained scores would also be determined Those factors that cannot be isolated and separately analyzed remain as "residual" factors

Analysis of variance as a method of estimating reliability is preferred by some psychologists, but it has not been widely used [9]

Reliability is also evaluated at times by means of statistical devices with which may be calculated consistency of performance from item to item within a test [10] This method introduces the assumption that the test is completely homogeneous as to functions measured, that is, that each item in the test measures precisely the same composite of mental functions as every other item In most tests this is a doubtful assumption, but if the assumption is warranted, the technique may be used

**Factors affecting the interpretation of reliability coefficients.** In addition to the considerations mentioned in connection with the several methods of estimating reliability, there are other factors that must be taken into account in interpreting reliability findings

*Range of ability* of the group tested affects the reliability coefficient If a reliability coefficient is found with a group that has a relatively small variation of the trait or function being measured, the coefficient will be relatively low If the group has a wider range in the trait or function, the coefficient will be higher (See Figure 1 3 ) Thus, a test having high reliability for a widely varying group does not necessarily have equal reliability for a significantly more homogeneous group of persons The reasons for this fact are several, one being the nature of the correlation process and the elements in the correlation formula

For illustrative purposes, suppose that we are dealing with a completely homogeneous group of individuals, with respect to one measure namely, chronological age Assume that everyone in the group is exactly ten years of age If they are an adequately representative

[9] Since analysis of variance as a method of estimating reliability requires more than knowledge of elementary statistics it will not be further elaborated here See R W B Jackson and G A Ferguson *Studies on The Reliability of Tests* Toronto Department of Educational Research University of Toronto, 1941 Also C Hoyt, "Test Reliability Obtained by Analysis of Variance" *Psychometrika* Vol 6 1941 pp 153–160

[10] See G F Kuder and M W Richardson, "The Theory of Estimation of Test Reliability," *Psychometrika* Vol 2, 1937, pp 151–160

sample of all ten year olds, the range in test score might be from extremely low to extremely high In this instance, since there is no deviation (or range) whatever in one of the measures (chronological age), the correlation coefficient for the two variables (test score and CA) will be zero [11] Such an extreme instance rarely occurs, but it does demonstrate that when there are possibilities for wide variations in one measure (in this instance, the test score) and very restricted possibilities in the other (in this instance, the CA) the coefficient is lowered

If the age range were two years instead of one, the coefficient of correlation would still be low, but not zero because *in general* the members of the older group tend to have higher test scores than do those in the younger But since there is a wide range of capacity within each group and overlapping of capacity between the two age groups the coefficient will be low

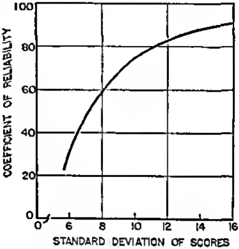



FIG 1 3 *Curve Showing Increase in Test Reliability as Variability of Group Increases*

A correlation coefficient reflects the *group* trends of the measures As persons increase in age mental capacity increases until maximum development is reached The correlation coefficient will reflect this fact But since there are wide differences in capacity within any age group and since there is considerable overlapping of capacity even among rather widely separated age groups, the coefficient will be affected by these facts also The result will be a coefficient lower than +1 00 (perfect correlation)

Thus, in correlation estimates of reliability, if the age range is wide,

---

[11] Inspection of the product moment correlation formula will show this to be the case

$$r = \frac{\Sigma(xy)}{N(SD_x SD_y)}$$

in which $\Sigma xy$ is the sum of products of the deviations of the paired scores $SD_x$ and $SD_y$ are the standard deviations of the two sets of measures

the group trends in scores (higher scores with higher ages) will have increased weight, as compared with a narrower age range in which the age trend has less weight In interpreting a reliability coefficient of a test, therefore, it is necessary to know the range of ages upon which the test was standardized

TABLE 1

*Raw Scores and Ranks of Students on Two Forms of an Arithmetic Test*

| | *Form X* | | *Form Y* | |
|---|---|---|---|---|
| *Student* | *Score* | *Rank* | *Score* | *Rank* |
| A | 90 | 1 | 88 | 2 |
| B | 87 | 2 | 89 | 1 |
| C | 83 | 3 | 76 | 5 |
| D | 78 | 4 | 77 | 4 |
| E | 72 | 5 | 80 | 3 |
| F | 70 | 6 | 65 | 7 |
| G | 68 | 7 | 64 | 8 |
| H | 65 | 8 | 67 | 6 |
| I | 60 | 9 | 53 | 10 |
| J | 54 | 10 | 57 | 9 |
| K | 51 | 11 | 49 | 11 |
| L | 47 | 12 | 45 | 14 |
| M | 46 | 13 | 48 | 12 |
| N | 43 | 14 | 47 | 13 |
| O | 39 | 15 | 44 | 15 |
| P | 38 | 16 | 42 | 16 |
| Q | 32 | 17 | 39 | 17 |
| R | 30 | 18 | 34 | 20 |
| S | 29 | 19 | 37 | 18 |
| T | 25 | 20 | 36 | 19 |

Just as in the foregoing illustrations, correlation coefficients were shown to be lowered by homogeneity in one of two variables so in estimating reliability the coefficient will be lowered by restricting the group s range of variation in the trait being measured An illustration will help to clarify this matter [12]

'In Table 1 are shown the raw scores and rankings of twenty students on two forms of an arithmetic test Looking at the two sets

[12] From *Test Service Bulletin* The Psychological Corporation No 44 May 1952

of rankings we see that changes in rank from one form to the other are minor, the ranks shift a little, but not importantly " A coefficient computed from these data is very high $r = .968$

'Now, however, let us examine only the rankings of the five top students Though for these five students the shifts in rank are the same as before, the importance of the shifts is greatly emphasized Whereas in the larger group Student C's change in rank from third to fifth represented only a 10 percent shift (two places out of twenty), his shift of two places in rank in the smaller top group is a 40 percent change (two places out of five) When the entire twenty represent the group on which we estimate the reliability of the arithmetic test going from third on form X to fifth on form Y still leaves the student as one of the best in this population If, on the other hand, reliability is being estimated only on the group consisting of the top five students, going from third to fifth means dropping from the middle to the bottom of this population—a radical change " A coefficient, if computed for just these five cases, is .50 (*rho*) [13]

'Note that it is not the smaller number of cases which brings about the lower coefficient It is the narrower range of talent which is responsible A coefficient based on five cases as widespread as the twenty (e g, Pupils A, E, J O, and T, who rank first, fifth, tenth fifteenth, and twentieth respectively on form X), would be at least as large as the coefficient based on all twenty students' [*rho* = +1 00]

Furthermore, when the variation among testees is narrow, the correlation between two sets of scores may also be lowered by chance factors and minor psychological factors Since individuals in such a group are closely clustered—that is, their true differences are small—the changes in scores and relative positions produced by extraneous factors are more significant than they would be in a widely divergent group

This illustration makes clear the fact that reliability coefficients of a given test may vary as the composition of the tested group changes, even though the performances of the testees themselves are unchanged Thus reliability data may show that a test discriminates satisfactorily over a *wide* range of the trait or capacity measured, but reliability may still be inadequate where *finer* and *more precise* dis-

[13] *Rho* represents the "rank-order" correlation coefficient It approximates closely the product moment coefficient (*r*)

criminations are necessary among individuals who vary within a narrow range

The practical significance of range and hence of ability is this in standardizing a test, its author must determine reliability with a group that is similar in average level of ability and in variation of scores to the group with whom the test is to be regularly used The user of a test should select an instrument that, among other things, provides reliability data based upon a sampling of persons who resemble closely the group of individuals he desires to test and study

*The time interval between testings* may be significant in the interpretations of reliability findings When reliability estimates are based upon odd-even correlations (internal consistency), or upon the scores of two equivalent forms of a test administered at a single sitting or within the same day, the results are uniformly affected by the examinees' physical condition and attitudes, and by the prevailing environmental conditions during the testing Such uniformity means that the factors external to the test itself are likely to affect both sets of test scores equally, thus increasing the degree of similarity of each person s two scores This condition tends to give higher reliability coefficients (that is gives higher estimates) regarding the instrument's predictive value than would be the case if the retest were given after a time interval

When there is a time interval the retest results will be affected by the normally expected fluctuations in individual performances and by changes in environmental conditions Thus while test results and reliability coefficients obtained at a *single sitting* or in a *single* day are most likely to estimate best the consistency of the instrument it self, they do not indicate stability of performance over a period of time as well as do coefficients obtained by the test retest method, using a time interval Conversely, the test retest method is the more likely to underestimate the internal consistency of a test because factors extraneous to it may affect the scores dissimilarly The extent to which the accuracy of a test is underestimated by the test retest method will depend upon the degree to which effective influencing conditions are inconsistent If the time interval has been quite long especially in the case of young children—perhaps three months or more—an individual s retest results may be influenced by peculiarities of his growth tempo or by other more or less enduring conditions such as emotional experiences, which may affect persons of any age

The longer the interval between tests, the more likely the lowering of the correlation due to intervening factors.

*The effects of practice and learning* during the interval will depend upon the content of the test being used and upon the examinee's experiences during the interval. For example, if some months have elapsed between two administrations of an educational achievement test, different pupils may have had different amounts and qualities of instruction during the period. The retest scores would, in part, reflect this instructional difference, thus the correlation coefficient would not be solely a reliability coefficient. Or, in the case of a personality test, individuals in therapy or after extensive counseling may have modified their attitudes, values, and behavior sufficiently to produce significant differences in test-retest results.

*Which method of estimating reliability is preferable* depends upon the problem at hand. Psychologists and educators are usually concerned with knowing (1) the internal consistency of a test, and (2) the predictive value of a test when it is subject only to the minor or accidental changes in conditions from day to day, rather than to fundamental changes resulting from permanent or semi-permanent changes effected by learning, developmental idiosyncracies, or disturbing emotional experiences. For the first purpose, the odd-even method, or the test-retest method, the tests being given at one sitting or within one day (using equivalent forms), is preferable. For the second purpose, the test-retest method, the tests being given within a week or two (using the same form or equivalent forms), is the preferable one. Under testing conditions that are not too different, the results of the second method will not be far removed from those obtained with the first. A test manual should provide information regarding internal consistency and test-retest results.

*Sub-test reliability is not always equal to total test reliability.* It has already been explained that, other factors being equal, the reliability of a test increases with increase in length, although not in direct proportion. This principle applies to those scales that consist of several different parts (called *sub tests*), each of which utilizes a different type of content. Nearly all group tests are of this kind, as are some of the individual scales (e.g., the Wechsler). For these instruments, the total test reliability is higher than that for each of the subtests. It is erroneous, therefore, to assume that the reliability coefficient for the whole may be applied to a part. For example, the Wechsler In-

telligence Scale for Children shows a full scale (nine sub-tests) reliability of .92 for a group of 200 children 7½ years of age, the coefficient having been calculated by means of the split-half technique.[14] Yet the reliabilities of the individual sub-tests, for the same group of children, ranged from a low of .59 to a high of .84. It is obvious that the total score is a more dependable index of the function or functions being measured than is any sub-test score.

*Consistency of scorers* is a factor in calculating test reliability. Some tests (such as the Stanford-Binet and, in particular, projective techniques) are not entirely objective in scoring, since the examiner at times finds it necessary to judge the correctness or quality of responses. For tests such as these it is necessary to know the extent of agreement in scoring found among two or more competent persons who have scored the same sets of responses. Test authors usually report such data in their manuals; and, in addition, other psychologists will have carried out and reported studies on this problem. Lack of agreement among scorers will adversely affect the reliability findings.

## VALIDITY

An index of validity shows the degree to which a test measures what it purports to measure, when compared with accepted criteria. The construction and use of a test imply that the instrument has been evaluated against accepted standards or other criteria which are regarded by experts as the best evidence of the traits or abilities to be measured by the test. Selection of satisfactory validation criteria and demonstration of an appropriate degree of validity is fundamental in psychological and educational testing.

The first necessary condition of a valid test is that it have an adequate degree of reliability. If the reliability coefficient of a test is zero, it cannot correlate with anything. A test that correlates poorly even with itself cannot correlate well with a measure of another variable.

**Operational and Functional Validity.** It is useful to recognize two kinds of validity, although they are not mutually exclusive. The first is known as *operational* validity; the second is *functional* validity.

By operational validity we simply mean that the tasks required by the test are adequate for the measurement and evaluation of certain

[14] Manual, New York: The Psychological Corporation, p. 13.

specified and defined psychological operations For example, the Seashore Measures of Musical Talent are actually tests of only *certain essential auditory aspects of musical talent,* but not of "musical talent" which psychologically involves much more than these auditory aspects Insofar as the Seashore tests differentiate correctly between persons in regard to the specified auditory processes, they are *operationally* valid On the other hand, these measures are *functionally* valid to the extent that they are efficient in predicting subsequent development of various degrees of skill and competence in the several aspects of music Thus, the functional validity of a test is the extent to which it is efficient in predicting and differentiating behavior or performance in a specified area under actual working and living conditions

Numerous other examples can be cited to illustrate the difference between functional and operational validity Thus, a peg board test (placing small metal pegs into a perforated board) may well measure manual and digital dexterity (operational), but it might be only slightly useful in predicting "mechanical ability" (functional) Again, a word and number checking test may be quite satisfactory as a measure of perception of details (operational), but it might have limited value in predicting "clerical ability"

It is obvious that functional validity is dependent, at least in part, upon the operational validity of the test The reason is that the psychological operations required by the test have been included because they have been found to be essential in certain actual situations in which testees will or might be placed Hence, if the psychological operations themselves are not measured with adequate validity, predictions of later performance will be adversely affected

**Criteria of Functional Validity.** The problems of selecting and utilizing satisfactory validating criteria vary with the several kinds of tests In constructing *tests of general ability* (intelligence), a common practice is to use some or all of the following scholastic marks, teachers' judgments of individuals' abilities, cumulative scholastic averages over a period of years, number of school grades completed, chronological age, and known groups The reasons for using these as criteria are (1) that scholastic records are evidence of mental ability even though influenced by factors other than intellectual ability, (2) that teachers are in a position to evaluate individual ability with some validity, because they observe their pupils over a long period and are able to

make inter-pupil comparisons, (3) that cumulative scholastic averages are more valid than marks or estimates of a single teacher because they represent combined judgments of performance over a longer period of time, (4) that on the whole the more able persons complete more formal education and reach higher levels in school and college, (5) that as individuals grow older their levels of intelligence increase until adult maximum is reached, (6) that definitely known groups, such as gifted, somewhat superior, mentally deficient, slow learning, and average groups will show differential performance on a valid test

The principal criteria in standardizing *tests of specific aptitudes* (e g , mechanical, musical) are marks in training courses and differentiation of known groups possessing the aptitude in varying degrees An example of known groups would be those working efficiently at each of several levels of a mechanical occupation and those in nonmechanical occupations It is highly desirable, of course, to use degree of success of actual performance in the vocation as an ultimate criterion

When the criterion of actual performance on the job is used, the following kinds of ratings are obtained ratings by supervisors, evaluation of the quality of the product, and rate of work However, the most frequently employed criteria for tests of specific aptitudes are marks and ratings in training courses, that is, criteria of capacity to learn the given skill or the profession, since aptitude tests are used largely to select individuals for training or education in the specified areas, although their use in employee selection is not inconsiderable

In *personnel work*, in business and industry, where specialized tests are used to select individuals for specific jobs, it is possible, indeed essential to use actual production records or performance ratings as criteria of test validity If, for example, a personnel department wants to know whether certain measures will identify the potentially best stenographers, the tests might be administered (1) to a group of employees of several quality levels to estimate the instruments differentiating efficiency, and (2) to newly employed personnel whose performance records, after an adequate period, would be correlated and otherwise analyzed against their test scores

*Tests of educational achievement* are validated against school marks and teachers' ratings Frequently, also, the criterion is "content validity" rather than some external standard. "Content validity"

means simply that the author of the test has determined upon its content by means of an analytical process and his own judgment, as well as that of experts in the subject-matter, as to what is appropriate and germane For example, in constructing a test of American history, the author examines what he believes to be representative textbooks, consults teachers of history, decides which topics are most significant and what their relative weights should be, and devises items he believes are most representative of these

*Tests of personality traits* present an especially difficult problem in validation Often the author of the test uses "face validity" Whenever an author bases his test upon his own analysis of what is to be evaluated or measured, without reference to prescribed content, as in textbooks, or without subjecting his device to comparison with other external standards, he is using "face validity" as his criterion At times, the traits presumably being measured by a particular test of personality have been included only by fiat The sounder tests in this category, however, are validated against actual behavior of the subjects and against clinical diagnoses But even these criteria present difficulties because they are themselves affected to an appreciable degree by the subjective judgments of the persons making the evaluations of behavior or the clinical diagnoses These matters will be discussed further in subsequent chapters

Criteria of validity may be *immediate, intermediate,* or *ultimate.* The use of marks in a particular course of study as a criterion in validating a test of specific aptitude is a case of an immediate criterion The cumulative average marks in an entire training curriculum can be regarded as the intermediate criterion, if performance on the job itself is the ultimate criterion If a test is being designed solely for the purpose of predicting marks in a single course (say, geometry), then those marks are the ultimate criterion Whether a criterion belongs in one or another of the categories depends upon the purpose of the test and the number of phases or steps that are available for use as criteria In fact, when validity findings on a test are to be interpreted, *the purposes for which the instrument is designed must be* taken into consideration A particular test may have different validities for different purposes, different age groups, different sex groups, etc The validity of a test lies in the correctness with which it measures at the time it is administered and in its predictive value for specified activities by specified groups

Factorial Validity. This method utilizes factor analysis techniques that are not within the scope of the present discussion Factor analysis theory, however, is discussed in Chapter 3, in connection with theories of intelligence Yet, since factorial validation is a method used with some tests, students should be familiar with the general nature of the theory [15]

Most tests of mental ability and of personality sample a composite of performances such as verbal knowledge and facility, number facility and quantitative reasoning, memory span, concept formation, etc Factor analysts maintain that these and others, especially when represented by a single composite index (such as mental age or intelligence quotient) are not 'functional unities" Analysts urge that they are not measures of a "pure" ability, that is, just one type of ability uncomplicated by others Thus, according to this theory, a test is said to have high *factorial validity* if it is a measure of one 'functional unity" (e g, word knowledge) to the exclusion of other elements as far as possible The factorial process aims to identify, by the method of intercorrelations and further statistical analysis, a list of 'functional unities' (also called 'primary mental abilities') within a test and the weight contributed by each of these to total performance on the test The ultimate goal is to devise tests each of which will measure only one 'functional unity" and be relatively independent of others (that is, show quite low intercorrelations) Such "pure' tests would then be used singly, or they might be used as subtests in a comprehensive measuring instrument, but even then each subtest is scored and rated independently for the purpose of obtaining a psychological profile for each person

If validation stops at factorial validity, 'operational' validity has been established But if, after factorial validity is established, we proceed to validate the test against criteria of later performance in working situations, we are making a "functional' validation The principal contribution of factorial validation is this instead of validating the total, undifferentiated instrument against functional criteria an effort is made to identify the component psychological elements and to establish their relative independence, and finally, to correlate these elements separately against functional criteria

Such analysis into psychological 'unities," or elements, is of value

---

[15] See J P Guilford "Factor Analysis in a Test Development Program" *Psychological Review* Vol 55, 1948, pp 79–94

when individuals are to be selected for specialized work or study and their performance predicted therein For example, since "mechanical aptitude" is not a simple, unitary skill, it is valuable to be able to identify which psychological elements bave most predictive value for a specified type of work Mechanical work may involve a high degree of spatial perception in one situation but not in another, or manual precision and speed, or comprehension of mechanical principles Also, higher than average intelligence is desirable for the practice, let us say, of both law and engineering In the former, word knowledge and verbal concept formation are the more significant, whereas in the latter spatial perception and quantitative reasoning are more significant than word knowledge Factorial analysis can assist in identifying the more limited and immediately relevant aspects of ability required in a given occupation or activity

**Face Validity.** This is a term that is used to characterize test materials which *appear* to measure that which the test author desires to measure Use of the term "validity" in such instances is hardly warranted, for the materials have not been objectively analyzed for validity In instances of face validity the author and those using the test simply assert in effect that the content of the test appears to he appropriate and to serve their purposes Face validity is found most often in personality inventories and in some of the more recently pub lished projective tests, notably the *Szondi* test (described in Chapter 19) This type of "validity" is also found at times in methods used in the selection of industrial personnel It is, however, unwarranted and should not be resorted to unless a relatively objective approach is impossible or not feasible

**Cross Validation.** This term refers to the process of validating a test by using a population sample other than the one on which the instrument was standardized The reason for using this method is that at times the original validity data may be spuriously high due to the operations of some chance factors that produce a higher correlation than is warranted As a matter of fact, however, once a test is put to use in a variety of situations and by many different persons, it is being constantly cross-validated, and if it does not prove to have high enough functional value, its use will be, or should be, discontinued

**Methods of Calculating Validity.** The most frequently used technique of estimating validity is the *simple correlation* of test scores with each

criterion A coefficient of a given magnitude cannot be arbitrarily specified as signifying or as not signifying validity Whenever a validity coefficient is positive and significant, it has some value In some instances, coefficients of only + 30 have been found useful Most coefficients, however, should be larger

ERROR SCORES ON *STORE PERSONNEL TEST FORM FS*

TRAINER RATINGS (Sum of learning ability, working speed, and overall fitness)

| | 56–54 | 53–51 | 50–48 | 47–45 | 44–42 | 41–39 | 38–36 | 35–33 | 32–30 | 29–27 | 26–24 | 23–21 | 20–18 | 17–15 | 14–12 | 11–9 | 8–6 | 5–3 | 2–0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 12 | | | | | | | | | | | | 2 | 2 | 1 | | 1 | 1 | | |
| 11 | | | | | | | | | | 1 | 2 | | 3 | 3 | 1 | 5 | 1 | | |
| 10 | | | | | | | | | | | 3 | 1 | 3 | 1 | 1 | 2 | 2 | | |
| 9 | | | | | | | | | 1 | | 5 | 4 | 3 | 6 | 4 | 3 | 1 | | |
| 8 | | | | | | | | | | 1 | 4 | 4 | 2 | 2 | 2 | 2 | 1 | | |
| 7 | | | | | | | | | | 3 | 4 | 3 | 1 | 2 | 2 | 4 | | | |
| 6 | | | | | | | 2 | | | 2 | 6 | 2 | 1 | | 2 | 1 | | | |
| 5 | | | | | | | | 1 | 1 | 4 | 3 | | 1 | 2 | 1 | 1 | | | |
| 4 | | | | | | | 1 | 1 | 3 | 1 | 1 | 1 | | | | | | | |
| 3 | | 1 | | | | 1 | 1 | 1 | | | 3 | | 1 | 3 | | | | | |
| 2 | | | | | | | | 1 | 1 | 1 | | 1 | | | | 1 | | | |
| 1 | | | | | | | | | | | | | 1 | | | | | | |

FIG 1 4 Chart for Pearson product moment correlation between number of errors made by 155 grocery store trainees on Part II of the experimental Store Personnel Test, Form FS and ratings made by the training staff $r = 46$

Figure 1 4 illustrates the simple correlation method [16] The test scores shown horizontally, were correlated with trainer ratings, shown vertically The number in each cell shows how many persons earned the scores of that cell as indicated on both axes For example, two persons who made between 21 and 23 errors on the test were given trainer ratings of 12, then going to the bottom of the same column we find that one person who also made between 21 and 23 errors had a trainer rating of 2 For this sampling of examinees, the

[16] From *Test Service Bulletin* No 37 1949 New York The Psychological Corporation The data of biserial and tetrachoric correlations that follow are also from this bulletin.

coefficient is 46, which is well within the range of validity coefficients most often found for a single criterion

The *biserial correlation coefficient* is used when one of the criteria is rated in terms of only two categories e g , "pass" or "fail", "satisfactory" or "unsatisfactory" The second measure, however, is given in terms of variable scores This method is used when the situation

TABLE 2

Biserial correlation between scores of 52 employed stenographers on the Seashore Bennett Stenographic Proficiency Test and their supervisors' ratings on stenographic ability $r_{bs} = 60$ (The Psychological Corporation)

| | Ratings on Stenographic Ability | | | |
|---|---|---|---|---|
| Test Scores | Below Average | Average | Above Average | Excellent |
| 19 | | 1 | 3 | 2 |
| 18 | | 3 | 3 | 5 |
| 17 | | 1 | 2 | 2 |
| 16 | | 1 | – | 2 |
| 15 | | 8 | 5 | |
| 14 | | 2 | – | |
| 13 | | 3 | 1 | |
| 12 | | 1 | 1 | |
| 11 | | – | – | |
| 10 | 2 | 2 | 1 | |
| 9 | – | – | | |
| 8 | 1 | 1 | | |
| Subtotals | 3 | 23 | 15 | 11 |
| Totals | 26 (Group 1) | | 26 (Group 2) | |

requires only a rough evaluation, as in the illustration presented in Table 2 Here we see that the four groupings on the basis of supervisors' ratings (below average, average, above average, excellent) have been reclassified into two categories (Groups 1 and 2) which have been correlated with stenographic proficiency test scores The biserial coefficient of 60 indicates that the proficiency test has considerable value in identifying stenographers who will function at satisfactory or highly satisfactory levels

The *tetrachoric coefficient of correlation* is an index that is found when a coarse classification of two measures is adequate for the pur-

pose at hand When this index is used, the ratings in *each* measure are grouped into only two classes, providing a "four fold' table The data in Table 2 have been so reclassified in Table 3, yielding a tetrachoric coefficient of + 60

Whether one uses the finer classifications necessary for calculating the product moment coefficient (the simple correlation above) or uses the coarse groupings shown in biserial and tetrachoric calculations will depend upon the nature of the data available and upon the purpose for which validation is to be used

TABLE 3

Four fold table for computation of tetrachoric correlation coefficient $r_{tet} = 60$ (*The Psychological Corporation*)

| Test Scores | Ratings 3–5 (Rated Low) | Ratings 6–8 (Rated High) | |
|---|---|---|---|
| 16–19 | 6 (11 5%) | 19 (36 5%) | High on Test |
| 8–15 | 17 (32 7%) | 10 (19 3%) | Low on Test |

*Multiple correlation* is a method whereby two or more criteria are statistically combined and correlated with the test score to yield a single coefficient Whereas the simple product moment coefficient indicates the degree of relationship (or co-variation) between two sets of measures, the multiple correlation coefficient shows the relationship between one set of measures (in this instance, test scores) and the composite of two or more other sets of measures (in this instance, the criteria) In other words, while a test might have a low or moderate correlation with a single criterion, it can have a quite significant correlation with several criteria taken together as a composite This is so because the several criteria in combination have more elements or factors in common with the test than does any one factor taken singly

*Expectancy tables* provide a relatively simple, straightforward, and very valuable method of estimating the predictive efficiency of a test.

The estimates are based upon the calculated probabilities that an individual who has a given test score will achieve a specified score or rating in the performance being predicted We might ask, as examples, the following questions What are the probabilities that a prospective college student scoring in the highest decile group on a ' scholastic aptitude ' (intelligence) test will remain in college a given number of terms? What are the probabilities that a child with an IQ of 80–85 will be able successfully to complete the work of the eighth grade? What are the probabilities that a candidate getting an average score on a stenographic proficiency test will achieve a rating of ' excellent" or "above average ' on the job? Appropriate expectancy tables are intended to answer these and similar questions

TABLE 4

*Decile Rank on a Scholastic Aptitude Test and Semesters Completed (in percents)*

| Decile Rank | Terms * 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| X | 98 | 95 | 94 | 90 | 89 | 88 | 88 |
| VII | 94 | 87 | 85 | 82 | 81 | 78 | 78 |
| VI | 92 | 82 | 79 | 74 | 74 | 73 | 73 |
| II | 85 | 71 | 66 | 61 | 59 | 57 | 56 |
| I | 81 | 67 | 60 | 52 | 51 | 49 | 48 |

* Decile rank X is the highest I is the lowest

Table 4 is an illustration in point It presents part of a larger table representing all ten decile groups

To take two items from Table 4, we may say the probability is that 88 in 100 of the students in the highest decile group on the scholastic aptitude test will complete their academic course, whereas only 48 in 100 of the lowest decile group will do so

Table 5 illustrates the use of expectancy data in personnel selection Inspection of this table shows that it may be used to indicate what percentage of individuals obtaining each of the several ratings on actually demonstrated stenographic ability may be found at each of the several levels on the proficiency test It is also possible to calculate the percentages by rows (instead of columns) so as to indicate the converse namely, the frequencies of the several ability ratings within each of the score intervals of the proficiency test

Comparison of Tables 4 and 5 demonstrates that expectancy tables need not be uniform The form and arrangement of data will depend upon the particular probabilities one desires to determine But all expectancy tables for tests have this in common they provide estimates of the probabilities that a certain level or quality of performance may be expected if the test score is known—that is, its functional validity in terms of probabilities in place of or, more often, in addition to a correlation coefficient

TABLE 5

*Expectancy table showing the number and percent of stenographers of various rated abilities who came from specified score groups on the* S B Stenographic Proficiency Test (N = 52, mean score = 15 4 S D = 2 9, r = 61, *score is average per letter for five letters* ) (The Psychological Corporation)

| *Number in each score group receiving each rating on stenographic ability* | | | | *Stenographic Proficiency Test Scores* | *Percent in each score group receiving each rating on stenographic ability* | | | |
|---|---|---|---|---|---|---|---|---|
| Below Average | Average | Above Average | Excellent | | Below Average | Average | Above Average | Excellent |
| | 4 | 6 | 7 | 18–19 | | 17 | 40 | 64 |
| | 2 | 2 | 4 | 16–17 | | 9 | 13 | 36 |
| | 10 | 5 | | 14–15 | | 44 | 33 | |
| | 4 | 1 | | 12–13 | | 17 | 7 | |
| 2 | 2 | 1 | | 10–11 | 67 | 9 | 7 | |
| 1 | 1 | | | 8–9 | 33 | 4 | | |
| 3 | 23 | 15 | 11 | | 100 | 100 | 100 | 100 |

*A cut-off score* is a special instance of the expectancy method It is a test score that is used as a point of demarcation between examinees who will be accepted and those who will be rejected For example, Table 6 shows several values from the Cornell Index (a personality inventory discussed in Chapter 17) that might be taken as cut-off scores

A low score on this inventory signifies fewer personality problems, hence it is more desirable The table reads If a cut-off score of 7 on the Cornell Index were used with this group of 1000 persons, 86 percent of those who were rejected after the interview would have been rejected also by the Index, but 28 percent of those accepted after the interview also would have been *rejected* by the index. The other per-

centages are read in the same way. Since the higher scores in this instance are the less desirable, and if 7 is taken as the cut-off level, we mean that a score of 7 or lower would be necessary for acceptance. If 23 were the cut-off, then anyone with that score or lower would be acceptable. In this instance the cut-off score becomes less selective as *it increases.*

TABLE 6

*Percent of Psychiatric Rejects * and Accepts * Identified by the Cornell Index (Reprinted by permission from the Manual. The Psychological Corporation.)*

| *Cut-off Level* | *400 Rejects* | *600 Accepts* |
|---|---|---|
| 7 | 86% | 28% |
| 13 | 74 | 13 |
| 23 | 50 | 4 |

* *Based upon psychiatric interviews*

If we are using a test on which larger scores signify higher and more desirable levels of the ability or trait being evaluated, then the cut-off score becomes the more selective as it is increased. Table 7 is a case in point.

TABLE 7

*Percent of Superior and Inferior Workers Identified by a Proficiency Test*

| | *Superior Workers* | | *Inferior Workers* | |
|---|---|---|---|---|
| Cut-off Score | *Accepted* | *Rejected* | *Accepted* | *Rejected* |
| 20 | 100% | 0% | 80% | 20% |
| 25 | 90 | 10 | 60 | 40 |
| 30 | 80 | 20 | 40 | 60 |

The hypothetical example shown in Table 7 is interpreted thus: If 20 were set as the minimum acceptable score, then all examinees who proved to be superior workers would have been employed; but so would 80 percent of those who proved to be inferior workers. The other cut-off scores are similarly interpreted

It is clear that cut-off scores are especially useful in instances where many more candidates are available than there are places to be filled, so that the cut-off level may be made highly selective, and where one is not concerned primarily with the individual candidates as such, but

rather with the places, jobs, or niches to be filled The purpose in using cut-off levels is to identify a maximum number of potentially superior or desirable persons and, at the same time, to eliminate a maximum number of inferior or undesirable individuals Since no test has perfect validity, and since the true potential of some persons may not be revealed by a single test, screening by means of cut-offs will not be perfect Some desirable persons will be rejected, whereas some undesirables will be selected Yet, cut-offs and other methods have a very considerable advantage over subjective procedures previously employed, for they provide the data for estimating with greatly increased objectivity and accuracy what are the chances of identifying the persons with the desired abilities or traits

*Other methods*, it will become apparent in later chapters, are used in addition to those already explained Among these others are, for example, the percent who are successful in adjacent age groups and in groups of known ability, significant increases in scores from age to age and from group to group, closeness of approximation of the distribution of scores to the normal frequency curve Also, in validating personality scales, extent of agreement by specialists in the interpretation of results is an accepted criterion

There are instances, too, when very low correlations are regarded as evidence of a test's validity For example, if one starts with the hypothesis that "mechanical aptitude" is a special ability and, as such, relatively independent of what is measured as "general intelligence," then in constructing a test of the former, one should, among other things, aim to devise a test which has a quite low or negligible correlation with the latter

**Item Analysis.** With very few exceptions, psychological tests (other than projective techniques) are made up of a large number of items The score on each item is added to the scores of the other items to obtain a subtest score or a total score, either or both of which are used in calculating reliability and validity Ultimately, however, the quality and merit of a test depend upon the individual items of which it is composed It is therefore necessary, in best practice, to analyze each item in the standardization process in order to retain only those that suit the purposes and rationale of the device being constructed Item analysis is thus an integral part of both the reliability and the validity of a test

In evaluating items, three aspects are, in the main, considered (1) the level of difficulty of each, (2) correlation of each item with the score of a subtest or with the total score of the whole test; (3) the degree to which each item differentiates between a high group and a low group (variously selected), or between several groups at different levels

The first of these aspects, *item difficulty,* is a matter of the percentage of individuals able to pass each item In practice, if an item is to distinguish between individuals, it should not be so easy that all persons can pass it, nor should it be so difficult that none are able to pass it [17] It can be demonstrated statistically that an item passed by 50 percent of a group discriminates between more pairs of persons than does an item passed by a smaller or larger group For example, if an item is attempted by 100 individuals and passed by only 10, and if the testees are taken by pairs, there are 900 (10 × 90) combinations in which that item can discriminate between paired members of that group If the item is passed by 50 in the group, then the number of possible discriminations between paired individuals is 2500 (50 × 50), this being the largest number possible, as the multiplication of any other proportions will show [18] Obviously not all items in a scale are or should be such as to be passed by 50 percent of the group Some are included that are passed by a large percentage and some by a small percentage, with many degrees between the extremes

There is no formula for determining the exact distribution of item difficulties A common practice is to select some items whose difficulty is at or close to the 50 percent level, and other items with a wide range of degree of difficulty, in terms of percent passing If all items selected for inclusion in a test were at the 50 percent level of difficulty, the

[17] Theoretically, it would be desirable that the test be so scaled that there is at least one item which can be passed by all for whom the test is intended For zero scores on a particular test do not necessarily mean absolute zero capacity in the function being measured, nor will all zero scores necessarily signify the same status Conversely, it would be desirable that a test be scaled upward to a level where no one for whom the test is intended is able to pass the highest item This aspect would require of course, that the test be constructed by a person superior to any of the intended testees

[18] It is not to be assumed that "50 percent passing" is necessarily the best criterion in placing an item in an age scale (like the Binet), as will be seen later

Percentages passing an item may be converted into scale values on the base line of the "normal curve, that is, into standard scores The assumption here is that the trait being tested by each item is distributed "normally" (bell shaped curve) in the population being tested

test would, theoretically, simply divide the testees into two groups namely, those above this predetermined dividing point and those below it Such items would not differentiate among the individuals in the group above the 50 percent level, nor among those below it Hence, for maximum differentiating efficiency, a test must contain items at various levels of difficulty as represented by percentages passing The final consideration will be the inclusion of items of such a range of difficulty as to yield the highest predictive value when compared with the criteria, taking into account the levels of the ability or trait to be measured and the degree of differentiation to be achieved

A second method of analyzing validity of individual items is to correlate each item against the score of the subtest of which it is a part (e g, information, arithmetical problems) to determine whether or not performance on it is consistent with performance on the subtest as a unit [19] This assumes, of course, that all items in the subtest are expected to be homogeneous, that is, that they measure the same psychological process or combination of processes

Each item may be correlated, also, against the score of the *whole* test In that case, the assumption is that all the items throughout the entire test are expected to be homogeneous in basic functions measured When an item is correlated against the subtest score, it is not necessarily expected to show a significant correlation with the whole-test score, because it may be the intention of the test's author to construct a scale whose subtest scores are relatively independent

The third technique is to analyze each item in respect to the performance of a low group and a high group, that is, low and high based upon scores on the test as a whole, or upon some external criteria whereby individuals can be classified As already stated, a very few items should be within the ability range of all, or nearly all, testees Others should be of increasing selectivity Some items should, of course, discriminate between two extreme groups, say, the highest and lowest 10 percent of the population tested, but it is desirable to have items whose selectivity extends beyond these narrow boundaries, items that would also dependably distinguish between, for example,

[19] The statistical method used for this purpose is the biserial correlation or the point biserial See any standard textbook on statistics For presentation of the problems and methods in item analyses see F B Davis *Item Analysis Data* Harvard Education Papers No 2 Graduate School of Education Harvard University, 1946 Also J A Long and P Sandiford, *The Validat on of Test Items* University of Toronto Press 1935

the highest one fourth and the second highest one-fourth, between the lowest one fourth and the next lowest one-fourth Kelley has offered evidence indicating that most marked and significant discrimination between extreme groups is obtained when item analysis is based upon the highest 27 percent and the lowest 27 percent of the group (The ratio of the obtained difference to the standard error is a maximum [20]) This method however, provides only a crude item differentiation, since it does not provide a basis for differentiating among the large middle group of the population, namely, about 50 percent

Using this method, one procedure would be to find what percentage of the highest 27 percent and what percentage of the lowest 27 percent passed each item, then, by statistical calculation, to determine if the difference between the two percentages is significant The same method can be followed with other proportions as well In fact, items may be analyzed with regard to a wide variety of group classifications Each item might, for example, be analyzed with reference to high, average, low average, and low groups classification being based upon total test scores or upon external validating criteria

**Validating Objectives.** The objective of all validating procedures is to make the most useful selection of test types and test items from among those available, such as to yield the highest prediction of the criterion or criteria *The first step in preparing such test items is insight into the psychological processes involved* The next prerequisite is that the items shall be well and precisely written Then, basically the ultimate decision as to what are the criteria of validity in any area of testing rests upon the analytical judgments of and agreement among those specialists best qualified to evaluate objectives and behavior, who take into account the purposes for which and the groups for whom the instrument is intended.

[20] T L Kelley The Selection of Upper and Lower Groups for the Validation of Items *Journal of Educational Psychology* Vol 30 1939 pp 17–24 For a detailed technical treatment of measurement techniques especially as related to personnel problems see R L Thorndike *Personnel Selection Test and Measurement Techniques* New York Wiley 1949

# 2.

## INTERPRETATION OF TEST SCORES: QUANTITATIVE AND QUALITATIVE

### AN INDEX OF RELATIVE RANK

The raw score (that is, the actual number of units or points) obtained by an individual on a test does not in itself have much if any, significance One test may yield a maximum score of 150, another 200, and a third 300 Obviously, then, any point score on one of these tests is not directly comparable with the same number of points on either of the others, a score of 43 on one test cannot be directly compared with a score of 43 on another Furthermore, the average scores of each of these will in all probability be different, as will the degree of variation of scores (called the *deviation*) both above and below the average For example, the average (*mean*) score of the first test for a given age is, let us say, 90, with approximately the middle two thirds of the scores falling between 75 and 105 For the second test the mean is, say, 120, with the middle two thirds of the scores between 100 and 140, while for the third test the mean is 180, with the range of the middle two thirds between 150 and 210

It is clear that if scores obtained on each of several tests are to be compared, indexes must be used which will express the relative significance of any given score, or what is known as *relative rank* In the example given above, assuming that all three tests are intended for the same group, the mean scores of 90, 120, and 180 have the same relative significance—that is, persons making these scores would be at the average in each Similarly, scores of 75, 100, and 150 have the same

relative significance in their respective tests, for persons getting these scores would be one standard deviation below the means (averages), which signifies that their scores surpass only about 16 percent of all the scores made by the population samplings upon whom the tests were standardized

Innumerable other comparable points and scores could be selected for illustration Obviously, however, such score-for-score comparisons would be extremely cumbersome and would, in each instance, have to be interpreted in terms of some common, meaningful index Hence, to facilitate interpretation, psychological tests (with few exceptions) provide tables of *age norms* or *grade norms*, or *percentile ranks*, or *decile ranks*, or *standard scores* These indexes are defined in the following paragraphs

**Norms.** A *norm* is the average or typical score on a particular test made by a specified population—for example, the average (mean) intelligence test score for ten year-olds, or twelve-year-olds, the mean score for fifth grade pupils on a test of arithmetic fundamentals Reference to a table of norms enables one to rank an individual s performance relative to his own or other age groups or grade groups Thus, for example, a child of ten may attain an intelligence test score that is average for his own age group, or for a population of nine-year-olds, or for those ten and a half, etc , or a fourth-grade pupil, on a test of arithmetic fundamentals, may score at the level typical for his grade, or for some grade above or below

**Mental Age.** By means of tables of norms, it is possible to assign an individual an "age" rating, on the basis of his performance on the particular test being used Thus, an individual, regardless of his age, who gets an intelligence test score that is equal to the norm of the ten-year-old population would have a "mental age" of ten as determined by that test If his score equaled the norm of the eleven-year-old population, his "mental age" would be eleven Hence we define *mental age* (MA) as the level of a person's mental ability as expressed in terms of the chronological age of average persons having the same level of mental ability

**Educational Age.** If a pupil obtains a total score on a battery of achievement tests (covering several school subjects) equal to the norm of pupils who are twelve years old, he is said to have an "educa-

tional age" (EA) of twelve *Educational age* is defined as the age equivalent of an individual's score on an achievement test as shown by age norms for the test in question, measuring achievement in a group of school subjects

In a similar manner, age levels can be determined for individuals on any test whose purpose is to rank persons on the basis of performance according to age There are times, however, when it is desirable to know the relative level at which a person is located, in respect only to a more narrowly specified group The group, for example, might be of a particular age or grade, adults at large, or a particular class of persons such as college freshmen For this purpose, the most commonly used indexes are percentile rank, decile rank, and standard scores

**Percentile Rank.**[1] An individual's percentile rank on a test designates the *percentage of cases or scores lying below it* Thus a person having a percentile rank of 20 ($P_{20}$) is situated above 20 percent of the group of which he is a member, or, otherwise stated, 20 percent of the group fall below this person's rank. A percentile rank of 70 ($P_{70}$) means that 70 percent fall below—and so on for any percentile rank in the scale In effect, this statistical device makes it possible to determine at which one hundredth part of the distribution of scores or cases any particular individual is located By this means a person's relative status, or position in the hierarchy, can be established with respect to the traits or functions being tested And, as will be seen, psychological measurement, unlike physical measurement, derives the greatest part of its significance from relative ranks ascribed to individuals rather than from units of measurement

**Decile Rank.** The decile rank is the same in principle as the percentile rank, but instead of designating the one-hundredth part of a distribution it designates the one tenth part in which any tested person is placed by his score The term *decile* technically means a dividing point By 'decile rank' we signify a range of scores between two dividing points Thus a testee who has a decile rank of 10 ($D_{10}$) is located in the highest 10 percent of the group, one whose decile rank is 9 ($D_9$) is in the second highest 10 percent, one whose decile rank is 1 ($D_1$) is in the lowest 10 percent of the group

[1] Also called "centile"

**Standard Score.** This index ($Z$) is somewhat less obvious in its meaning than percentile and decile ranks, although it, too, designates the individual s position with respect to the total range and distribution of scores The standard score indicates how far, in terms of standard deviation, a particular score is removed from the mean of the distribution The mean is taken as the zero point, and standard scores are given as plus or minus If the distributions of scores of two or more tests are approximately normal ('bell shaped'), standard scores derived from one distribution may be compared with those derived from the others

The formula is

$$Z = \frac{X - M}{SD}$$

in which $X$ is an individual score $M$ is the mean of the distribution and $SD$ its standard deviation

Assume, for example, that the mean IQ of a group is 100 and that the standard deviation is 14 In this distribution an individual reaching an IQ of 114 has a $Z$ score of $+ 1\ 0$ Another individual having an IQ of 79 has a $Z$ score of $- 1\ 5$

TABLE 8

Proportions of Cases, or Area Under the Curve, Corresponding to Given Standard Scores

| Z Score | Approximate Percent of Cases from the Mean |
|---|---|
| 25 | 10 |
| 50 | 19 |
| 75 | 27 |
| 1 00 | 34 |
| 1 25 | 39 |
| 1 75 | 46 |
| 2 00 | 48 |
| 3 00 | 49 8 |

Standard scores must ultimately be given percentile values to express their full significance Since the number of cases encompassed within a given number of standard deviations in a normal distribution is mathematically fixed it is always possible to translate a $Z$ score into a percentile value Thus a person having a $Z$ score of $+ 1\ 0$ has a percentile rank of approximately 84, that is, his score surpasses 84

tional age" (EA) of twelve *Educational age* is defined as the age equivalent of an individual's score on an achievement test as shown by age norms for the test in question, measuring achievement in a group of school subjects

In a similar manner, age levels can be determined for individuals on any test whose purpose is to rank persons on the basis of performance according to age There are times, however, when it is desirable to know the relative level at which a person is located, in respect only to a more narrowly specified group The group, for example, might be of a particular age or grade, adults at large, or a particular class of persons such as college freshmen For this purpose, the most commonly used indexes are percentile rank, decile rank, and standard scores

**Percentile Rank.**[1] An individual's percentile rank on a test designates the percentage of cases or scores lying below it Thus a person having a percentile rank of 20 ($P_{20}$) is situated above 20 percent of the group of which he is a member, or, otherwise stated, 20 percent of the group fall below this person's rank A percentile rank of 70 ($P_{70}$) means that 70 percent fall below—and so on for any percentile rank in the scale In effect, this statistical device makes it possible to determine at which one-hundredth part of the distribution of scores or cases any particular individual is located By this means a person's relative status, or position in the hierarchy, can be established with respect to the traits or functions being tested And, as will be seen, psychological measurement, unlike physical measurement, derives the greatest part of its significance from relative ranks ascribed to individuals rather than from units of measurement

**Decile Rank.** The decile rank is the same in principle as the percentile rank, but instead of designating the one-hundredth part of a distribution, it designates the one-tenth part in which any tested person is placed by his score The term *decile* technically means a dividing point By "decile rank" we signify a range of scores between two dividing points Thus a testee who has a decile rank of 10 ($D_{10}$) is located in the highest 10 percent of the group, one whose decile rank is 9 ($D_9$) is in the second highest 10 percent, one whose decile rank is 1 ($D_1$) is in the lowest 10 percent of the group

[1] Also called "centile"

**Standard Score.** This index ($Z$) is somewhat less obvious in its meaning than percentile and decile ranks, although it, too, designates the individual s position with respect to the total range and distribution of scores The standard score indicates how far, in terms of standard deviation, a particular score is removed from the mean of the distribution The mean is taken as the zero point, and standard scores are given as plus or minus If the distributions of scores of two or more tests are approximately normal ("bell-shaped'), standard scores derived from one distribution may be compared with those derived from the others

The formula is

$$Z = \frac{X - M}{SD}$$

in which $X$ is an individual score, $M$ is the mean of the distribution and $SD$ its standard deviation

Assume, for example, that the mean IQ of a group is 100 and that the standard deviation is 14 In this distribution an individual reaching an IQ of 114 has a $Z$ score of $+ 1\ 0$ Another individual having an IQ of 79 has a $Z$ score of $- 1\ 5$

TABLE 8

Proportions of Cases, or Area Under the Curve, Corresponding to Given Standard Scores

| *Z Score* | *Approximate Percent of Cases from the Mean* |
|---|---|
| 25 | 10 |
| 50 | 19 |
| 75 | 27 |
| 1 00 | 34 |
| 1 25 | 39 |
| 1 75 | 46 |
| 2 00 | 48 |
| 3 00 | 49 8 |

Standard scores must ultimately be given percentile values to express their full significance Since the number of cases encompassed within a given number of standard deviations in a normal distribution is mathematically fixed, it is always possible to translate a $Z$ score into a percentile value Thus a person having a $Z$ score of $+ 1\ 0$ has a percentile rank of approximately 84, that is, his score surpasses 84

percent of the scores in the group The person having a Z score of $-1.5$ has a percentile rank of approximately 7, surpassing only 7 percent of the scores Table 8 below shows several standard scores and their percentile values, for illustrative purposes

As an index of relative rank, the standard score is preferred by some psychologists because it is a well-defined property of the normal curve, representing a fixed and uniform number of units throughout the scale Percentiles and deciles, on the other hand, are positions of rank in a group and do not represent equal units of individual differences

**Quotients.** The use of tables of norms for the determination of the several kinds of test "ages" has already been mentioned [2] Now, in addition to these performance "ages," it is a rather common practice to determine a "quotient" Of these, the most widely known is the *intelligence quotient* (IQ), found by the simple formula

$$IQ = \frac{MA}{CA}(100)$$

in which MA is the individual's mental age and CA is his chronological age Thus, it is clear that the IQ is the ratio of one's mental age to his life or chronological age (multiplied by 100 to remove the decimal) and indicates *rate of mental development* or *degree of brightness* If mental development keeps pace with life age,[3] the quotient is 100, if mental development lags or is accelerated, the quotient will be less than or greater than 100, depending upon the degree of retardation or acceleration

Tests of educational achievement, as already stated, yield educational ages (EA) This index may be divided by CA to give an *educational quotient* (EQ), that is, an index showing, presumably, whether a person's knowledge and understanding of a group of school subjects are commensurate with his life age, or whether above or below what would be expected of him for his age

Educational age may be divided also by MA instead of CA If that is done, the index is the *accomplishment or achievement quotient* (AQ) The reason for using the mental age instead of the chronologi-

[2] There are a few tests for which mental ages are not derived from tables of norms notable among these being the Stanford Binet Scale and the Merrill Palmer Scale These tests and their scoring technique are presented in subsequent chapters

[3] That is until maximum capacity is reached.

cal age in the denominator is that the former is regarded as the more valid index of learning capacity. Hence, dividing the EA by the MA yields a quotient which shows whether or not the individual is working up to mental capacity as revealed by the intelligence test.[4]

Subsequently, more will have to be said concerning these and other quotients. For the present, however, it is to be noted that quotients, like the other rating devices already presented, are in fact indexes whose significance is to be found to a considerable extent in the relative status they give individuals.

Let us assume that we are dealing with three boys, all of the same age. Suppose that their intelligence quotients are 50, 100, and 150. Since these are numerical ratios $\left(\frac{MA}{CA} \times 100\right)$, it is natural to assume that they have a *quantitative* significance. So they do—for they indicate rate of mental development. But these quotients also have a *qualitative* significance—for, among other things, they indicate each boy's position in the "hierarchy of intelligence." If the measure of intelligence is valid, the boy having the IQ of 50 is seriously retarded and is in the lowest one percent of the population in respect to the psychological functions being tested, the boy with the IQ of 100 is the "typical" or "average" individual, midway up (or down) in the distribution of intelligence, and the boy having the IQ of 150 is very superior and belongs in the top percentile rank of the group.

Qualitative significance of the intelligence quotient can be illustrated further by asking this question: Is the brightest of these three boys one and one-half times as intelligent as the "average" boy, and three times as intelligent as the retarded one? The fact is that this question cannot be answered in terms of numbers; it is impossible actually to say how many "times" more capable or less capable one is than the others, because *the IQ is not a percent*. But each of these quotients has certain connotations. In this example, the qualified school or clinical psychologist will be able to draw important inferences from each boy's IQ regarding rate and quality of school learning, extent and level of educability, vocational possibilities and levels, and probable types of interests.

The boy with an IQ of 50 probably will not be able to complete

[4] There are several theoretical problems concerning the AQ (such as the logical fallacy of an AQ greater than 100) which will not be dealt with at this stage, but will be treated in Chapter 14.

more than the second grade, the boy having the IQ of 100 should be able to complete twelve grades, the boy with an IQ of 150 will be able to progress in education as far as his interests and motives indicate Obviously, too, the kinds of occupations that will be open to the first boy are very limited, those open to the second will be numerous, those open to the third will be practically unrestricted so far as the factor of mental capacity is concerned And the same may be said of the range of interests in general that will be within the scope of each These facts are of clinical significance, but at present there are no psychological or statistical means whereby one can calculate how many times more or less capable one person is as compared with another

A caution is necessary at this point The inferences drawn in the preceding paragraph cannot be based solely upon the numerical IQ value without reference to the clinical features in the test performances or other factors not shown by the numerical index We have assumed that there are no complicating factors and that the IQ s are valid measures of the capacities and performances of the three boys The boy with 150 IQ however, might be an unstable personality who is failing in most or all of his school subjects The boy with 100 IQ might have been penalized on the test by linguistic factors And the boy with an IQ of 50 might show a ' scatter (inconsistency and variation) of performance indicating emotional disturbance rather than intellectual impoverishment Occasionally, also it will be found that a high test rating may be attributable to an inconsistently high level of performance on one or a few types of subtests (e g memory span word knowledge), just as conversely, it occasionally happens that a person s IQ is depressed by an inconsistently low performance on one or a few subtests [5]

## PSYCHOLOGICAL MEASUREMENT CONTRASTED WITH PHYSICAL MEASUREMENT

The indexes thus far presented—ages percentile rank decile rank standard score, and quotients—are intended to emphasize the principle that basically all psychological tests yield results that rank individuals *in relation to* their fellows

[5] By inconsistent" we mean that the individual s levels of performance on these few subtests differ markedly in one direction or the other from the general and more uniform levels of his scores on the other subtests

The raw scores obtained on psychological tests are not comparable with or similar to the values obtained in the measurement of physical traits or phenomena, as, for example in measuring length, weight, or light intensity. In the physical realm, the units of measurement are fixed and constant throughout the entire scale. An inch, a pound, a candle-power—each has the same value and physical significance at whatever place on the scale it is measured Psychological measurement, by contrast, is more difficult and is confronted by special problems. In the first place, it is not possible to determine the "inherent" difficulty of an item in a psychological test in terms of constant units, as it is possible to find the length or weight of any object. Whereas in the measurement of physical phenomena it can be found that a given object is of *X* length and *Y* weight, and hence, let us say, twice as long and three times as heavy as another object with which it is being compared, no such *direct* measurement and comparison are possible in psychological testing. In this realm, the measurement value of a test item is dependent basically upon the percentage of persons able to pass the item in the population group for whom the test is intended.

If in the testing of a particular ability, one item is passed by only ten percent of a group, whereas another item is passed by fifty percent, it cannot be said that the first is five times as difficult as the second, because "percent passing" is not a unit in the sense that an inch or a pound is. What can be said is that an individual able to deal successfully with the first item belongs in the highest decile group, while one who cannot pass the first but is able to pass the second falls at the midpoint or "average" level of the group in respect to that item of the test. This interpretation is significant psychologically and educationally. Or, to use another instance, in the case of the Stanford-Binet Scale the age level at which an item is placed—and hence its value in the scale—is determined by the age group in which the "average" individuals pass that item. Thus, anyone able to solve a reasoning problem placed at the ten-year level, for example, but failing to pass reasoning problems at the eleven-year or higher levels, may be said to have typical ten-year-old ability in respect to that mental task.

This leads directly into the problem of the meaning of "mental age" and other "age" units. *Mental age* may be defined as the level of a person's mental development expressed in terms of the *chronological age* of average individuals of the same level of mental development.

Or, otherwise stated, mental age is an index showing one's level of mental development, corresponding to the level of mental development of average persons of the coinciding *chronological age* Thus, if a child s mental age is teo, he has reached the level of mental development attained by average children of ten years, regardless of the actual life age of the child being tested

Now, suppose there are four individuals having mental ages, respectively, of five, six, twelve, and thirteen Is the difference between

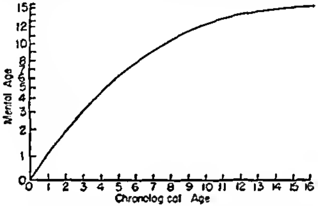



FIG 2 1 *Hypothetical Curve of Mental Growth,* Illustrating Decreasing Yearly Increments

mental ages five and six the same as that between mental ages twelve and thirteen? It is not, for, as measured by mental tests, the rate of mental development at five and six years of age is more rapid than subsequently, hence, the increment between the earlier years is greater Figure 2 1 is the curve accepted by most psychologists as representing rate of mental development The outstanding feature of that curve, for the problem under consideration, is the fact that it rises at a *decreasing rate* with increasing age The curve is "negatively accelerated" Or, psychologically speaking, with each succeeding year, until maximum development is attained, the amount of increment in mental growth and development is less than in the preceding year

It is obvious, therefore, that each successive year of mental age added to an individual s level represents something less in measurable growth and development than the preceding year's increment In other words, mental age units are not uniform, they simply rank an indi-

vidual with respect to the average mental capacity of an age group The same principle—nonuniformity of age units—applies also to all other types of psychological tests which translate their scores into age equivalents

Although psychological testing would be facilitated and would be given a higher degree of precision if its measuring units were fixed and uniform, the fact is, nevertheless, that the available indexes of *relative* rank are essential educationally and clinically in evaluating the status of an individual's mental development, his educational progress, his particular aptitudes, his social maturity, and even certain nonintellective aspects of personality Experimentally, too, these same indexes have been indispensable in studying a host of practical and theoretical problems such as sex differences, effects of environmental conditions, inheritance of intellectual capacity and of special aptitudes, occupational differentiation, racial differences, relationships between physical and mental development, and many others

## CLINICAL ASPECTS

Scores, whether raw or converted, do not suffice for the complete interpretation of an individual's performances on psychological tests The several aspects of test standardization thus far presented are concerned with the performance of *groups* of persons and with *average relationships* revealed by statistical treatment of results It happens, however, that while certain types of test items meet some or most of the statistical requirements of validity, they are unsatisfactory as indicators of intelligence when used for clinical purposes For example, on the Stanford-Binet scale, the percentage of adults able to repeat eight digits forward (*digit span* test) is approximately the same as the percentage who can solve one of the more difficult reasoning problems Yet, in clinical examinations psychologists find some adult mental defectives who can pass the former test, though it never happens that a true mental defective can succeed with the latter What this means, in effect, is that statistical validation of a test item is not always sufficient, it must be supplemented by the pragmatic criterion of use with a wide variety of individuals in a variety of situations to show whether or not it has discriminative value as between individuals at the several levels of ability

Psychological tests, as already noted, are standardized on the basis of the performance of a representative population, and an individual s

Or, otherwise stated, mental age is an index showing one's level of mental development, corresponding to the level of mental development of average persons of the coinciding *chronological age* Thus, if a child's mental age is ten, he has reached the level of mental development attained by average childreo of ten years, regardless of the actual life age of the child being tested

Now, suppose there are four individuals having mental ages, respectively, of five, six, twelve, and thirteen Is the difference between

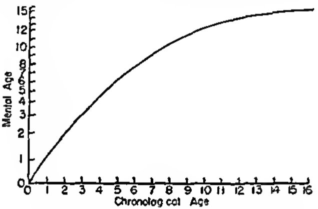



FIG 2 1 *Hypothetical Curve of Mental Growth, Illustrating Decreasing Yearly Increments*

mental ages five and six the same as that between mental ages twelve and thirteen? It is not, for, as measured by meotal tests, the rate of mental development at five aod six years of age is more rapid thao subsequently, hence, the increment between the earlier years is greater Figure 2 1 is the curve accepted by most psychologists as representing rate of mental development The outstaoding feature of that curve, for the problem uoder consideration, is the fact that it rises at a *decreasing rate* with increasing age The curve is "negatively accelerated" Or, psychologically speaking, with each succeeding year, uotil maximum development is attained, the amount of incremeot in mental growth and development is less than in the preceding year

It is obvious, therefore, that each successive year of mental age added to an individual's level represents something less in measurable growth and development than the preceding year's increment In other words, mental age units are not uniform, they simply rank an indi-

vidual with respect to the average mental capacity of an age group The same principle—nonuniformity of age units—applies also to all other types of psychological tests which translate their scores into age equivalents

Although psychological testing would be facilitated and would be given a higher degree of precision if its measuring units were fixed and uniform, the fact is, nevertheless, that the available indexes of *relative* rank are essential educationally and clinically in evaluating the status of an individual's mental development, his educational progress, his particular aptitudes, his social maturity, and even certain nonintellective aspects of personality Experimentally, too, these same indexes have been indispensable in studying a host of practical and theoretical problems such as sex differences, effects of environmental conditions, inheritance of intellectual capacity and of special aptitudes, occupational differentiation, racial differences, relationships between physical and mental development, and many others

## CLINICAL ASPECTS

Scores, whether raw or converted, do not suffice for the complete interpretation of an individual s performances on psychological tests The several aspects of test standardization thus far presented are concerned with the performance of *groups* of persons and with *average relationships* revealed by statistical treatment of results It happens, however, that while certain types of test items meet some or most of the statistical requirements of validity, they are unsatisfactory as indicators of intelligence when used for clinical purposes For example, on the Stanford Binet scale, the percentage of adults able to repeat eight digits forward (*digit span* test) is approximately the same as the percentage who can solve one of the more difficult reasoning problems Yet, in clinical examinations psychologists find some adult mental defectives who can pass the former test, though it never happens that a true mental defective can succeed with the latter What this means, in effect, is that statistical validation of a test item is not always sufficient, it must be supplemented by the pragmatic criterion of use with a wide variety of individuals in a variety of situations to show whether or not it has discriminative value as between individuals at the several levels of ability

Psychological tests, as already noted, are standardized on the basis of the performance of a representative population, and an individual's

rating is determined by the relationship of his performance to that of a group as a whole Thus we have the several "ages" (e g, mental age) and "quotients" (e g, intelligence quotient), percentile and decile ranks, and standard scores Any useful test should yield one or more of these ratings In more recent years, however, without denying the usefulness and value of these indexes of relative status, increasing emphasis has been placed upon "patterns" of performance as clinical aids to psychological diagnosis and counseling

A person's responses to tests are now frequently analyzed for the purpose of discovering whether he shows any special abilities or disabilities, whether there are marked discrepancies between responses on some types of materials as against responses on others, whether certain psychological processes seem to be impaired or are markedly superior to others within the individual A general contrast, for example, might be found between tests involving verbal materials and those which are nonverbal in character, the associative processes might be disturbed, memory or spatial perception might be found to deviate markedly in one direction or another from an individual's general level of capacity Recent investigations have indicated that "patterns' of response may be useful in differentiating and diagnosing the several categories of maladjusted and abnormal personalities, as well as for discerning more clearly the mental defectives

Also, it has been found that persons of equivalent general mental status may have different "patterns' of performance, or abilities, which in sum, nevertheless, give them much the same over-all and general ratings in terms of a single index (e g, mental age, percentile rank) That is to say, it is possible for two persons to have test ratings which are *numerically* similar but whose "mental organizations" are dissimilar, since the components of each total rating differ to a greater or lesser degree from those of the other

If, therefore, the psychologist's concern is not primarily with group trends or averages but rather with a particular individual, he will, to be sure, want to know the age level of performance and the consequent "quotient', but he will also analyze the details of the individual's performance for the purpose of discovering that person's particular pattern or idiom, in order to discern his particular form of mental organization, specific evidences of retardation, or disability, if any, and details of his development

In more recent years there has been a partial shift in emphasis from

almost exclusive concern with the analysis of abilities and methods of psychological measurement, as such, to an examination of individual performance and individual idiom, and to the individual as a functioning and dynamic unit After all, any given test measures only a segment of a total personality, that segment is an integral part of the totality and is influenced by the whole Hence, the psychologist who is concerned with insight into the nature of an individual s abilities must be able to *evaluate* a person's performance as well as *measure* it The data and indexes derived from psychological tests are, for the most part, objectively determined, but their clinical use involves judgment, subjective assessment and interpretation, based upon a variety of data from several sources The experienced clinical examiner will supplement the test's numerical results with his observations of the testee's attitudes during the examination and the manner in which he attacks the problems of the test his degree of confidence or dependence, his cooperativeness or apathy, his negativism or resentment, the richness or paucity of his responses The individual test situation can thus be, in effect, an occasion for general psychological observations—really a penetrating psychological interview

Ability not only to score a test but to assess and interpret responses and to evaluate the individual s behavior during the examination is a clinical art that the psychologist develops from working with persons rather than with tests alone, though for the practice of his art he must, of course, thoroughly understand the psychological and statistical foundations and hypotheses upon which the tests are based

A few specific instances of the *qualitative* analysis and interpretation of test responses will illustrate the kinds of observations that constitute the clinical aspects that supplement numerical scoring

> Word definitions are generally acceptable at a fairly elementary level but they vary in level and quality from purely concrete, to functional, to conceptual or abstract Differences in quality level are indicative of differences in modes of thinking It happens at times, also that some words are emotionally charged for the examinee in which case his definition and behavioral response may be revealing

> Some test items permit the exercise of considerable freedom in response These responses may reveal the examinee s attitudes values, and modes of meeting life situations In this category are test items that ask, 'What is the thing to do when ?" Or, 'Why should we ?' The subject s reactions to such items the qualities of his verbalizations in making the responses, and the presence or absence

of strong feelings reveal some of the nonintellective aspects of his personality

The subject's specific comments while performing a task are of possible significance in regard to his attitude toward himself, or toward an authority figure (the examiner), or toward other individuals and institutions in his environment

Responses to items or random comments may reveal hostilities and anxieties, or wholesome cooperativeness and security

The manner of speech, the use of expletives, halting and fumbling restless movements, blushing, or, on the other hand, ease of speech a relaxed attitude, mild criticism of one s own performance provide valuable clues to the testee's personality

Character disorders may be indicated by impetuous and uncritical responses that are incorrect but are given with assurance and pretentiousness

The subject's ability to direct his attention toward, to concentrate upon, and to organize a task are often revealed by his mode of approach to a test problem

The selective character, if any, of a person's vocabulary and information (two subtests widely used) will shed light upon his experiences, interests, cultural background

A personality trait such as compulsiveness (as opposed to desirable thoroughness and self criticism) may be revealed by excessively detailed responses and by numerous and unnecessary alternative responses

Some types of responses indicate pathological or psychotic states for example, erroneously bizarre responses by an otherwise intelligent person (e g, London is in Africa, the population of the United States is 1,500,000), by disjointed and irrelevant responses, and by distorted interpretations of the task or problem

Detection of organic damage through selected kinds of subtests, e g, disturbance of the visual motor function as indicated by the diamond copying test (Stanford Binet) and the object assembly test (Bellevue), among others

Scatter analysis (discussed in detail in Chapter 15) is essential to the discernment of superior, inferior, and impaired psychological functions

Sensitized observations on the part of the examiner will enable him, in general, to evaluate *how* the subject proceeded in both success and failure [6]

---

[6] Illustrations of these qualitative interpretations will be given in Chapter 14, on clinical aspects

The findings on a test—whether it is of general intelligence, specific aptitude, personality, or school learning—indicate the present status of each person examined They do not, however, tell the psychological examiner by what course the person arrived there, nor do they indicate specifically what factors were operative in his development The clinical approach, while accepting and utilizing standardized tests and norms, insists upon viewing and evaluating any given individual's performance and status in the light of a variety of other measurements, observations, and activities of that individual, and upon interpreting the objective quantitative data according to the part they have in the total For instance, children who have suffered from prolonged and serious nutritional disturbances or deficiencies or who are suffering from severe anemia will appear listless, apathetic, and deficient in mental capacity according to standardized tests Other children, of apparently retarded mental development, may be suffering from serious deficiencies of vitamin B complex, while still others may measure at a level of retardation because of emotional pressures and "blocking"[7] Furthermore, the performance of some children on standardized tests is of inferior quantity and quality because they developed under conditions of psychological impoverishment, whereas test norms are based on the assumption that all individuals being examined have had approximately equal opportunity, in the grosser aspects of environment, for mental development Often, of course, that is not the case These facts, and others of the same kind, indicate that in the case of some individuals, performance and consequent relative status may be impaired by nutritional deficiencies, by emotional handicaps, or by other unfavorable environmental conditions

The fact that psychologists are tending to put increasing emphasis upon the significance of the individual pattern in test performance (especially in diagnosing cases of behavior and educational maladjustment) and upon the individual as a whole does not mean that statistical and group studies are unnecessary Such studies are essential in providing norms against which any individual's performance may be projected in the process of evaluation, in giving more precise meaning and significance to any single score, in demonstrating the great range of human variability in any trait or function, and in providing the

---

[7] A condition in which the functioning of mental abilities is impeded due to the individual's emotional state or mental conflict

means of more precise study of interrelationships among psychological traits and functions

## DIFFERENCE BETWEEN NORMS AND STANDARDS

Norms, as already explained, are average scores or values determined by actual measurement of a group of persons who are representative of a specified population, for example, all twelve year-old boys, all fourth-grade children, all native-born male adults Norms, therefore, are averages *obtained under prevailing conditions*—good, bad, or indifferent These norms may well " reflect all the sins of omission or commission in their [the people s] nurture and must be critically examined lest we set up as desirable norms for achievement what are but accidental outcomes of our unsystematic and unenlightened nurture of children "[*] In other words, a norm of psychological performance or of a physical trait is not necessarily one with which we should be satisfied, for it reflects development under conditions that may be and often are much less than optimal As an example, consider the "average vocabulary' of the eighth-grade pupil The average, or norm, of this group will be dependent in part upon their opportunities from earliest childhood for the acquisition and use of language Their opportunities might have been extremely poor, moderately satisfactory, very good, or at any other level between these three Norms of performance in respect to some of the psychological processes measured by means of tests of intelligence and of specific aptitude are likewise dependent upon conditions and opportunities present during the course of development Norms of height, weight, and other body measurements will also reflect past conditions of nutrition and health

It is necessary, therefore, to distinguish between norms, on the one hand, and standards, on the other, for a standard is the *desired* goal or objective, which may well be above the obtained norm and can be achieved only under improved conditions of development It is possible that the grade norms for reading rate and comprehension are below what they could be under improved educational conditions and teaching methods, that age and grade norms for numerical ability are below what they might be, that universal nursery school and kinder-

[*] L. K Frank "Research in Child Psychology History and Prospect" in *Child Behavior and Development* R G Barker et al editors New York McGraw Hill Book Co 1943 p 9

garten experience would promote children's perceptions of form and color and would improve their motor skills, and hence raise the norms, that universal optimal nutrition would raise age norms for height and weight etc

Psychological tests measure traits and functions as they exist under present conditions They do not provide the psychologist and educator with an index of what *ought* to be, except by implication and insofar as obtained results might raise certain suspicions, doubts, and queries in the minds of investigators

## FACTORS IN SELECTING A TEST

By way of summary, the following factors are given as those to be considered in selecting a psychological test

**Norms.** The test must provide appropriate and accurate norms, whether they be in the form of age, grade, percentile rank, standard score, or any other type Norms should be meaningful with regard to the purposes for which the test is intended and to the groups of persons with whom it is to be used

**Administering and Scoring** The procedures of administering the test should be objective and the test items should be amenable to relatively objective and simple scoring insofar as the nature of the instrument permits Individually administered tests, like the Stanford Binet at times require insightful judgment in scoring responses Interpretation and evaluation of responses are even more significant in the scoring and analysis of projective techniques for assessing personality

**Time Requirements** The length of the test should not be so great as to produce boredom, satiation, or negativism, for when these set in, the subject does not perform at his best level Specific time limits cannot be prescribed for all tests or for all types of testees, but in general, shorter time requirements are indicated for younger children and for the mentally retarded In the case of both of these groups of subjects, the attention span is relatively brief, hence, it may be necessary at times to complete an examination in two sessions

***Interest Level*** *Test items should be of sufficient interest to motivate* the individuals for whom they are intended Particular items and types of problems devised to measure given functions must be suitable to the

age levels of examinees Thus in constructing an intelligence test for the entire range of adult capacity from very low to very high it is necessary that the items placed even at the very low levels should be of the sort that will interest an adult rather than a child even though these low level adults may be inferior to some children so far as mental capacity is concerned

The Popolation Sample The manual of a test should state in detail the nature of the population sample on which the instrument was standardized and upon which norms are based The information given should include the following total number of cases age range and number at each age level, number of each sex, geographic distribution socio-economic status and number in each category For some tests it will be relevaot indeed oecessary to have informatioo on some of the following school grade distribution, number of years of schooling completed, amount and kind of special training (especially for tests of specific aptitudes) special or "abnormal" adjustment problems and history (especially for tests of personality) In short, the prospective user of a test must be certain that the test has been standardized on an appropriate sample of the populatioo and *for the same or similar purposes* as those contemplated by the prospective user This principle seems axiomatic yet it is not always given due consideration

The Functions or Traits Measured. The test manual should not only state the purpose of the instrument it should also provide so far as possible an analysis (psychological and statistical) of the functions or traits being measured

Reliabilities Coefficients of reliability should be provided not only for total scores but for part scores as well wherever possible Also reliability coefficients are desirable for each of the several age levels and ability levels included within the range of the test Furthermore the manual should state which method or methods have been used in calculating the test s reliability Here the prospective user of a test must look for information that will help him answer the question "Reliable for whom and for what purposes?"

Validity Data on validity are of several kinds in addition to coefficients of correlation e g expectancy tables Known groups significance of differences between age levels etc The test manual should

explain the characteristics of the criterion *groups*, the nature of other criteria used, the validity of the total test, and the validity of the subtests It is desirable, also, to have data regarding validity at each of the several age and ability levels Here again, an answer must be sought to the question 'Valid for whom and for what purposes?'

**Reports of Experiments** An ideal to be aspired to is to have test manuals (subsequent to the first or earliest editions) include summaries, findings, and interpretations of the most important experimental studies to which the test has been subjected by various psychologists Such information will help users to understand more fully the nature of the test and the factors affecting performance on it, thus making for sounder interpretation of results obtained by those who use it For example What is the influence of cultural factors? Of practice? Of time limits? Of psychotherapy or counseling?

Psychological tests are scientifically constructed instruments, based upon psychological and statistical principles, as explained in the preceding pages Familiarity with these principles should provide students with a sounder comprehension of both the values and the limitations of tests than they would obtain from using and interpreting these instruments in a mechanical manner It is also true that when we test human subjects we are dealing with individuals who do not behave like mechanisms under complete control, with all environmental forces likewise under control and measurable On the contrary, human behavior is often subtle and the psychological forces motivating or influencing persons in a test situation may be elusive and difficult of evaluation Furthermore, as has been indicated, since the quantitative data of psychological testing are not as definite, precise, and uniform as are the data of physical measurements, the interpretation of test findings is the more difficult For these reasons, we have emphasized not only the well-defined scientific principles and procedures of testing but we have stressed the *qualitative* and *clinical* aspects which are essential, if test findings are to be of the greatest value to the individuals examined

age levels of examinees Thus, in constructing an intelligence test for the entire range of adult capacity, from very low to very high, it is necessary that the items placed even at the very low levels should be of the sort that will interest an adult rather than a child, even though these low level adults may be inferior to some children so far as mental capacity is concerned

**The Population Sample.** The manual of a test should state in detail the nature of the population sample on which the instrument was standardized and upon which norms are based The information given should include the following total number of cases, age range and number at each age level, number of each sex geographic distribution socio-economic status and number in each category For some tests it will be relevant, indeed necessary, to have information on some of the following school grade distribution, number of years of schooling completed, amount and kind of special training (especially for tests of specific aptitudes), special or "abnormal" adjustment problems and history (especially for tests of personality) In short, the prospective user of a test must be certain that the test has been standardized on an appropriate sample of the population and for the same or similar purposes as those contemplated by the prospective user This principle seems axiomatic, yet it is not always given due consideration

**The Functions or Traits Measured** The test manual should not only state the purpose of the instrument, it should also provide, so far as possible, an analysis (psychological and statistical) of the functions or traits being measured

**Reliabilities** Coefficients of reliability should be provided not only for total scores but for part scores as well, wherever possible Also reliability coefficients are desirable for each of the several age levels and ability levels included within the range of the test Furthermore, the manual should state which method or methods have been used in calculating the test s reliability Here, the prospective user of a test must look for information that will help him answer the question "Reliable for whom and for what purposes?"

**Validity.** Data on validity are of several kinds, in addition to coefficients of correlation, e g, expectancy tables known groups, significance of differences between age levels etc The test manual should

explain the characteristics of the criterion *groups*, the nature of other criteria used, the validity of the total test, and the validity of the subtests It is desirable, also, to have data regarding validity at each of the several age and ability levels Here again, an answer must be sought to the question "Valid for whom and for what purposes?"

**Reports of Experiments.** An ideal to be aspired to is to have test manuals (subsequent to the first or earliest editions) include summaries, findings, and interpretations of the most important experimental studies to which the test has been subjected by various psychologists Such information will help users to understand more fully the nature of the test and the factors affecting performance on it, thus making for sounder interpretation of results obtained by those who use it For example What is the influence of cultural factors? Of practice? Of time limits? Of psychotherapy or counseling?

Psychological tests are scientifically constructed instruments, based upon psychological and statistical principles, as explained in the preceding pages Familiarity with these principles should provide students with a sounder comprehension of both the values and the limitations of tests than they would obtain from using and interpreting these instruments in a mechanical manner It is also true that when we test human subjects we are dealing with individuals who do not behave like mechanisms under complete control, with all environmental forces likewise under control and measurable On the contrary, human behavior is often subtle and the psychological forces motivating or influencing persons in a test situation may be elusive and difficult of evaluation Furthermore, as has been indicated, since the quantitative data of psychological testing are not as definite, precise, and uniform as are the data of physical measurements, the interpretation of test findings is the more difficult For these reasons, we have emphasized not only the well-defined scientific principles and procedures of testing, but we have stressed the *qualitative* and *clinical* aspects which are essential, if test findings are to be of the greatest value to the individuals examined

# 3.

## DEFINITIONS AND ANALYSES OF INTELLIGENCE

### DEFINITIONS OF INTELLIGENCE

If intelligence is to be measured and assessed, it is necessary to define it, at least tentatively. A variety of definitions have been given by psychologists; but, as a matter of fact, each of them can be classified into one of several groups.

One group of definitions places the emphasis upon *adjustment or adaptation of the individual to his total environment*, or to limited aspects thereof. According to definitions of this type, intelligence is general mental adaptability to new problems and new situations of life; or, otherwise stated, it is the capacity to reorganize one's behavior patterns so as to act more effectively and more appropriately in novel situations Thus, the more intelligent person would be one who can more easily and more extensively vary his behavior as changing conditions demand; he has numerous possible responses and is capable of greater creative reorganization of behavior, whereas the less intelligent person has fewer responses and is less creative. The more intelligent person, accordingly, can deal with a greater number and a greater variety of situations than the less intelligent; he is able to encompass a wider field and to expand his area of activity beyond that of the less intelligent.

A second type of definition states that intelligence is the *ability to learn*. According to this definition, then, a person's intelligence is a matter of the extent to which he is educable, in the broadest sense. The

more intelligent the individual is, the more readily and extensively is he able to learn, hence, also, the greater is his possible range of experience and activity.

Still others have defined intelligence as the *ability to carry on abstract thinking*. This means the effective use of concepts and symbols in dealing with situations, especially those presenting a problem to be solved through the use of verbal and numerical symbols. Binet's conception of intelligence belongs largely in this category, for he maintained that it is the capacity to reason well, to judge well, and to be self critical.

It should be apparent that the three foregoing categories of definitions are not and cannot be mutually exclusive. For the most part, their authors differ in emphasis. Obviously, ability to learn must provide the foundation for adjustment and adaptation to changing or new conditions. And a person may be expected to have learned more or less from situations he had encountered and to which he had made adjustments previously. For if this were not the case, he would have to start anew in every situation which confronted him; there would be no difference between the behavior of an experienced person and that of a novice.

There are, of course, individual differences in respect to learning capacity and in ability to retain, interpret, organize and apply what has been learned; thus previous experiences will have different significance and different learning value for different persons. And it is learning capacity that constitutes the basis of adjustment and adaptation, although, as will become apparent in later chapters, important nonintellective factors affect adjustment and adaptation.

Yet learning capacity, in the sense only of acquisition of information and knowledge, is not a sufficient criterion by which to evaluate a person's intelligence. Psychologists and laymen alike are agreed that a person who can reorganize and apply what he has acquired for the purpose of dealing with varied and novel situations is more intelligent than one who is capable of little beyond repeating what he had previously acquired, or than one whose behavior follows stereotyped patterns without insight into the essential elements and relations of each new situation. Thus a definition of intelligence as the capacity to behave appropriately and effectively in new situations and a definition of intelligence as the ability to learn are in fact two aspects of the same process.

The third type of definition is also inseparable from the other two A person *learns* abstractions—principally verbal and numerical —through experience, through contact with and perception of the objects, events, qualities, relationships, etc, for which the symbols stand Thus, the word "dog" has meaning for a child because it has come to represent a class of objects with which he has become familiar The word "green" represents a quality he has perceived as an aspect of a variety of objects The word "charity," for the individual who has developed sufficiently to understand the concept, has a certain connotation by virtue of the fact that he has experienced events which have been labeled as charitable The number "five" is meaning ful to a person when, as a result of experience with concrete objects, he apprehends the term as representing not only ordinal position but summation as well Furthermore, if it is to be said that an individual has fully learned to deal with the symbols of abstraction, then it must be true that he understands that the *word is not the thing* or *the quality* for which it stands He understands that words and numbers are abstractions which represent objects, events, qualities, relations, etc, but which, in thinking, can be dealt with *as if* they were the things themselves This aspect of intelligence—the ability to use symbols—is itself the result of an individual's development and learning And in its turn, the mastery and utilization of symbols promotes further learning—for it is hardly necessary to labor the point that without language and number, the range of one's learning would be seriously restricted

Ability to carry on abstract thinking, it is easy to see, contributes to a person's ability to adjust or adapt to changing or new situations, because through the use of symbols we are enabled to think through a problem without spending time and effort on sheer trial and error in action, we are enabled to marshal, evaluate, and deal with past experiences, and we are enabled to project our thinking forward In other words, through the use of symbols and abstract thinking, man is able considerably to enlarge his range of behaving and adjusting, to extend his horizons, and to transcend the immediate concrete and specific situation

## TWO COMPREHENSIVE DEFINITIONS

In more recent years, two definitions of intelligence have appeared which, in effect, combine and extend the three types of defi-

nitions already presented One writer states [1] 'Intelligence is the aggregate or global capacity of the individual to act purposefully, to think rationally and to deal effectively with his environment' The reader can readily compare this defioition with those already presented and analyze it with a view to discerning similarities and differences It will be ooted, of course, that this definition encompasses the other three Although it does not specifically meotion learning ability, yet that is surely implied Two new aspects, however, are added The definition specifically states that an individual s intelligence is revealed by his behavior as a whole ("global"), and that intelligence involves behavior toward a goal, which may be more or less immediate ('purposefully") A third aspect is presented by the author in his elaboration of the definition, namely that "drive" and 'incentive" enter into intelligent behavior This aspect is probably included and implied in capacity 'to act purposefully" and "to deal effectively' with one s environment, as stated in the definition

The inclusion of 'drive," 'incentive," and the like as aspects of intelligence is of very doubtful validity, for to do so is to confuse the issue, the testing instrument, and the results obtained It is true, of course, that effective utilization of a person's intelligence depends upon the extent and degree to which he employs it Nevertheless, a single testing device which attempts to combine the measurement of intellectual with nonintellectual traits, without providing for differentiation between the two, would not succeed adequately in either respect [2] This is not to say that in assessing an individual s intelligence and personality as a "whole" we should ignore 'drive," incentive 'interest," etc, for the competent psychological examiner will and must evaluate these and other nonintellectual traits in presenting his test results Furthermore, as will be seen in later chapters of this book special psychological instruments are available for the evaluation of nonintellectual traits of personality which the clinician may use to supplement results of intelligence tests if he believes they are necessary

[1] D Wechsler *The Measurement of Adult Intelligence* Baltimore Williams and Wilkins 1944 p 3 Actually the usefulness of Wechsler s Bellevue scale is not dependent upon his conception of intelligence This scale does not in fact incorporate all aspects of the definition

[2] The Rorschach test attempts to measure both nonintellectual and intellectual traits of personality This test is discussed in Chapter 19

Stoddard [3] offers the following definition "Intelligence is the ability to undertake activities that are characterized by (1) difficulty, (2) complexity, (3) abstractness, (4) economy, (5) adaptiveness to a goal, (6) social value, and (7) the emergence of originals, and to maintain such activities under conditions that demand a concentration of energy and a resistance to emotional forces " Here again, the reader will note that this definition does in fact include the first three types of definitions presented, but it goes beyond these in several respects The author specifies the several attributes of intelligence, and in his enumeration are several not included in earlier definitions

Degree or level of "difficulty" is implied in all definitions, but Stoddard's contribution here lies in the fact that he rightly insists we must, in testing distinguish between true differences in degree of difficulty and differences that only seem to exist, as between two or more test items, whereas, in fact, there are no inherent differences in difficulty For example, the accumulation of pieces of rare information and the ability to define unusual words are not in themselves true measures of difficulty, they may reflect only differences in experience On the other hand, however, over and above disparity in experiences between various age groups, true differences in difficulty do exist between *problems* that can be solved, let us say, by a group of "average" ten-year-old children and those which can be solved by an "average" group of eight year-olds or nine-year-olds

'Complexity" refers to the number of different kinds and varieties of tasks that can be successfully dealt with According to this attribute of intelligence, the individual who is able to deal successfully with several different kinds of tasks, at a given level of difficulty, is more intelligent than a person who can successfully undertake fewer kinds of tasks at the same level of difficulty "Complexity," however, means not simply the addition of one type of performance to others, on the contrary, it means the capacity to assimilate new abilities, to integrate them with others, and thus to reorganize one's patterns or forms of intelligent behavior

"Abstractness"—that is, operating with symbols, especially at levels of analysis and interpretation—has already been discussed For Stoddard, this attribute "lies at the heart of intelligence as defined"

'Economy' refers to the rate at which mental tasks are performed

[3] G D Stoddard *The Meaning of Intelligence* New York Macmillan, 1943 p 4

and problems solved Assuming that the problems are solved equally well, that the solutions are equally effective, the individual working more rapidly would be regarded as the more able, according to this attribute Acceptance of "economy" as an attribute of intelligence means that tests would impose time limits which should differentiate among individuals in respect to their rates of performance of tasks and solutions of problems at given levels of difficulty and degrees of complexity

"Adaptiveness to a goal" implies an approach that is more than aimlessly meeting and solving new situations as they arise This attribute means that intelligent action is directed toward a goal or a purpose The more comprehensive the goal, the larger and more complete the purpose, the more is intelligent action required

The student, after examining representative tests of intelligence, might well question whether they do, or even could, satisfactorily test this last attribute, or whether the problems and tasks included in the tests are rather oversimplified and segmental examples of problems and courses of action that a person has to confront and deal with in actual life situations If the test items are of the latter kind, then their value and validity as measures of intelligence must be shown by the fact that they do indeed predict to an adequate degree the manner and effectiveness with which the testee will deal with and solve actual life situations of broader scope In other words, what are the predictive values of the items, tasks, and problems included in a test?

The inclusion of "social value" as an attribute of intelligence is of doubtful validity, and debatable at best, for this criterion is essentially moral or ethical, or a matter of subjective evaluation The basis of "social value" is group acceptability If this attribute were applied in evaluating intelligence, we should have to minimize our estimates of the intelligence of individuals whose thinking and solutions of problems are not necessarily consistent with accepted social forms, though they might be "ahead of their time", and of individuals who are capable of difficult and complex mental operations, but whose mental activities lead to no apparent or demonstrable practical and social values While we may well value more highly the individual whose mental operations culminate in desirable, acceptable, and useful social outcomes, the inclusion of "social value" as an attribute of intelligence would confuse attempts to measure the other, and valid, attributes by injecting largely subjective conceptions of what is socially acceptable

or unacceptable or indifferent It will be seen later that "social value" is hardly present in current tests of intelligence, although, of course, some psychologists, like Stoddard, take the position that it should be

"The emergence of originals" as an attribute of intelligence is the ability to create something new and different, it is a characteristic of a high order of thinking and of individuals at the superior end of the distribution of intelligence Examples of this attribute in operation are the development of a new scientific principle, the discovery of unique relationships in observed data or phenomena, the development of a new machine-design, the development of a new technological process, the new organization and new interpretation of historical or social facts, a creatively original painting or musical composition It is undoubtedly true that current tests of intelligence provide little opportunity for the measurement of emergence of originals The question, then, so far as these tests are concerned, is whether this attribute, *creative originality, is really dependent upon a combination of abilities* which are actually measured by available tests and whether the results obtained by means of the tests are indicative of the degree to which a person possesses this attribute Some psychologists, Stoddard among them, believe that at present the tests of intelligence do not satisfactorily discern and rate an individual's intellectual originality Others maintain that the hierarchy of abilities established by means of the tests enables us to identify the persons who possess originality in greater or lesser degree [4]

Stoddard's last two conditions of intelligent behavior—"concentration of energy" and "resistance to emotional forces"—are subject to the same criticism as Wechsler's inclusion of "drive" and "incentive" Motivation and ability to exert sustained effort are usually regarded as nonintellectual aspects of activity and are certainly recognized as playing highly important roles in one's general effectiveness But to introduce them into a test of mental ability would be to confuse and probably to invalidate efforts to arrive at a reasonably valid measure of the level of intelligent activity at which a given person is *able to*

[4] The observation of psychologists is that people who show creative originality almost without exception, score high or very high on intelligence tests but many persons may score very high on these tests without having exceptional powers of originality We need tests of originality, but in view of the very nature of the concept and its expressions, such tests cannot very well be standardized.

*operate* regardless of whether he actually does operate at that level in all situations

While tests of intelligence do not directly measure motivation and concentration of energy on the solution of problems, the psychological examiner does in fact try to develop or encourage conditions wherein the persons being examined will operate at their maximum levels of ability This can be more nearly achieved in administering an individual test than in administering group tests Furthermore, if an individual is not adequately motivated, is not expending a maximum of energy during the test, or is handicapped by emotional factors during the testing, these conditions can be discerned much more readily during the examination of one person at a time than during the examination of a group all at once Indeed, during group testing there may be instances of persons whose test results are vitiated by the effects of the unfavorable conditions mentioned without that fact being known to the examiner This possibility is a disadvantage of group testing

Of course, no single test or series ("battery") of tests can provide an unfailing index or a guarantee of motivation, energy output, or freedom from emotional blocking in all future situations requiring intelligent behavior For man is not a static being, nor does the environment in which he lives remain static Long term motives and immediate incentives will change, values and interests will change The affective (emotional) quality of a person's experiences will influence his subsequent behavior, including situations requiring the utilization of intelligence Tests of intelligence now in use are not intended to determine the extent to which an individual will in the future concentrate his energy on problems demanding the use of his intelligence, nor to determine whether it is probable that he will be able to remain free from emotional blockings A variety of personality rating scales and inventories and projective techniques have been devised to evaluate these nonintellectual traits Although tests of intelligence can and will no doubt be improved so that greater demands will be made upon concentration of attention and sustained effort than is the case with most tests at present, psychologists believe they are warranted in assuming that a qualified examiner will be able to determine whether or not a given person's performance on an *individual* test represents his maximum level at that time They believe, also, that most persons can be motivated to perform at their best levels when taking a *group* test

In any testing of groups this must be reasonably well assured, as well as assumed [5]

Although, as stated intelligence tests are not designed to measure a person s emotional status and other nonintellectual aspects of person ality, there is at present a trend, especially among clinical psychologists, to analyze test performance for evidence of emotional states, personality 'dynamics" and for 'differential diagnosis' (that is, for symptoms of neuroses psychoses, or other atypical states) This trend, and the basic idea, while still in a rather early experimental stage, are not without merit and some validity This aspect of test interpretation which demands sensitive clinical insights, will be discussed in a later chapter under scatter analysis and under projective analyses of test behavior

## IMPLICATIONS FOR TEST DESIGN AND CONTENT

Definitions of intelligence are of more than theoretical importance The conception of intelligence which a psychologist holds will affect, to some extent at least the content of the test he develops Yet, at the same time, an examination of a representative group of tests reveals the fact that although some are different from others in certain aspects, they all, nevertheless do have much in common It would be incorrect to say, for instance, that certain tests exemplify exclusively the definition that intelligence is the capacity to learn The fact, then, that psychologists emerge with tests having considerable similarity, though they might start with different definitions must mean that their definitions differ largely in respect to emphases and that, as already pointed out they are interdependent

Early experimenters in mental testing attempted to measure general intellectual capacity by means of a single type of test which measured only a single capacity, usually a sensory process or association, or attention Thus they identified general intellectual capacity with a

[5] The "catch" lies however in the fact that in any large group of persons being tested it is not unlikely that there will be a few at least, who are not adequately motivated or who are handicapped by emotional difficulties Herein is a source of error in group measurement. The discovery of these nonmotivated or blocked individuals will depend upon whether or not each one s rating on the test is scrutinized and evaluated in the light of other and perhaps conflicting evidence Where there is reason to believe that group-test performance is spuriously low in the case of a given individual it is desirable to re-examine using an individual rather than a group test.

single function Their efforts were fruitless Later, however, experiment showed that a *variety of test materials* yielded more accurate and more useful results when validated against accepted criteria of intelligent activity Psychologists, in seeking to encompass a greater variety of items in their tests, and thus to produce more useful and successful instruments, regardless of the exact definition with which each one started out, found that inevitably their testing instruments were broader than their definitions Current tests thus have more than a little in common in spite of differences in details of their content Inspection of their content will show that in varying degrees they are measures of some aspects of learning in what is assumed to be a reasonably uniform environment for all persons [6] that novel situations and problems are presented, and that ability to carry on abstract thinking is tested through the utilization of symbols and ideas In spection will demonstrate also that most of the tests fail to meet the more comprehensive and long term attributes suggested in Stoddard s definition

So far as available tests are concerned it is important to bear in mind that the French psychologist Binet (the 'father of modern mental testing) took the position that it made little difference what specific tasks and items were incorporated into a test provided that in some degree each part was a measure of the individual s general capacity Whether or not this condition is met will depend of course, upon the definition of intelligence regarded as most adequate by the designer of a test and upon the *criteria of intelligent activity against* which test results are checked for validity Developments have been such that in spite of some differences in definition and in spite of some differences in external appearances psychologists believe that their tests are reasonably sound because they are related to and have value in predicting the likelihood of intelligent activity in life situations

## THREE "KINDS" OF INTELLIGENCE

Some psychologists believe that several kinds of intelligence should be distinguished from one another Noteworthy among them is E L Thorndike who has divided intelligent activity into three types namely, (1) social intelligence, or ability to understand and deal with persons, (2) concrete intelligence, or ability to understand and deal

[6] This raises the much-debated problem of heredity and environment as factors in the development of intelligence

with things, as in skilled trades and in working with the appliances of science, (3) abstract intelligence, or ability to understand and deal with verbal and mathematical symbols

The merit of this classification of types of intelligent activity, for psychological testing and diagnosis, lies in the fact that it indicates several realms in which persons might be functioning and implies that separate and sufficiently specialized tests might be devised to measure how effectively persons are functioning in each

While it is true that in the case of any given person the scores attained on a test of ability to deal with verbal and numerical abstractions might differ appreciably from those attained by him on a test of social relationships and insights, or on one of "concrete" intelligence it is true, nevertheless that when a *representative group* of individuals are tested, the correlations between the types of tests are found to be positive and significant, both statistically and psychologically For example, correlations between tests of verbal and of concrete abilities vary from about 25 to about 45, the average being about 30–35 While this is a somewhat low average correlation, it still indicates that some communality of function is being measured This index, being so far from unity, also means that there are numerous individuals whose relative scores do not correspond closely or whose two relative scores may be discrepant This fact points up the important psychological principle that the data and status of any single person may be inconsistent with the *general trend* Study of the individual, and the ways in which and the reasons why he deviates from or exemplifies general trends, is one concern of the clinical psychologist

Of the three kinds of abilities enumerated above, abstract intelligence is the one that receives greatest weight and is most pronounced in current tests of intelligence—that is, whenever the test is designed for use with persons who are presumed to have reached a level where they reasonably may be expected to have developed facility in dealing with concepts and symbols

Even tests that present the subject with things" rather than with ideas and symbols are not devoid of demands upon ability to conceptualize and make abstractions, although testees need not necessarily state these in the form of language and number For example, when a subject is required to arrange a series of pictures into a sequential and meaningful whole, he must at some stage form a concept of "the whole' if his response is to be correct by some other means than pure

chance He must do this, also, in assembling parts into an integrated unit (called "object assembly") The same is true of the child who is asked which is the "prettiest" of two pictures ("aesthetic comparison"), for he must have a concept of "prettiness," however unarticulated it may be There are in use many other types of test items that deal with things but still require more or less ability in concept formation Among these are—to name a few—object classification, tracing the shorter of two routes in a maze, identifying objects by use, supplying missing parts in the drawing of a human figure, etc In short, the fact that some types of test items do not employ language or number does not necessarily signify that they make no demands upon ability to reason at a level of concept formation and abstraction

It is true that at the earliest developmental levels there are tasks that depend upon visual-motor skill, such as tying a bow knot, grasping a ring, holding a pencil and scribbling, manipulating cubes, and the like These types of tests, however, are restricted principally to the first eighteen months of life They are useful as developmental indicators, but they have only slight predictive value for later development of mental abilities, as measured by tests at more advanced levels [7]

The role of ability to deal with ideas and symbols (words and numbers) as a measure of concept formation and abstraction is of increasing importance in tests of general ability (intelligence) as age level increases Proportions of verbal and numerical tests, on the one hand, and nonverbal, nonnumerical, on the other, undergo change at different age levels, [8] some tests include a larger proportion of the latter than do others, even at the adolescent and adult levels [9] These differences are not haphazard, nor are they matters of individual whim, they depend upon the purposes of the test and the test author's conception of intelligence and its constituent parts It will be seen in later chapters that the correlations between various tests of general ability are quite marked—and at times high or very high—thus indicating that to an appreciable degree in these tests the verbal and

[7] See Chapter 8 'Scales for Infants and Preschool Children,' in which this problem is discussed at length

[8] Cf the Revised Stanford Binet Scale below the age 5 level Also the Merrill Palmer Scale, the Minnesota Preschool Test the Detroit Kindergarten Test etc

[9] Cf the Dearborn Group tests the Otis tests the Kuhlmann Anderson tests, the Wechsler Bellevue Intelligence Tests the Pintner Paterson scale

numerical items on the one hand and the nonverbal, nonnumerical on the other are measuring the same or closely related functions [10] High intercorrelations do not always mean that the same functions are being measured by the tests concerned, such correlations may reflect other common factors which affect the tests being correlated This, however, is quite improbable as an explanation of test intercorrelations

## ANALYSES OF MENTAL ABILITY

The definitions of intelligence thus far discussed are functional in character, that is, they state how intelligence operates through learning, adaptation, abstract thinking But, in addition, psychologists have been concerned to know the fundamental nature and structure of intelligence They have made analyses in an effort to determine its underlying *factors* Or, otherwise stated, the purpose of these analyses has been to discover, if possible, the elements, or components of intelligence not only for a better theoretical understanding of this complex process, but also to learn what might be the implications for the design and construction of mental tests

It is not to be inferred, however, that the dynamics of intelligent activity can be adequately understood merely by enumerating and characterizing the components, whatever they might be Whatever the components, they do not operate independently or in isolation Understanding the dynamic aspects of mental activity requires some means of characterizing the organization of factors, their interrelationships, and their relation to motivational forces

Essentially, the experimental method followed is this a rather large number of separate tests more or less diverse in character, are given to an adequate sampling of the population The results of each type of test are correlated with those of all the others The coefficients of correlation are then subjected to various techniques of statistical analysis in an effort to discover the extent of common ground between them (technically known as *communality*), and their degree of independence These statistical methods are known as *factor analysis* [11] The particular theory or structure of intelligence educed from the statistical operations will depend upon the expert's own interpretation of

[10] The statements in this paragraph do not mean that the nonverbal materials in the tests are measures of mechanical ability Generally they are believed by the authors of the tests to measure the same psychological processes as do the verbal materials but by means of different content.

[11] To be discussed later in this chapter

the analysis,[12] and the experts differ in their interpretations These differences, however, need not invalidate the use of well-standardized psychological tests, for, as will be seen, theoretical differences thus far have not had far-reaching consequences as regards the kinds of intelligence tests constructed

**The Multi-factor Theory.** Thorndike's multi-factor theory of intelligence is at one extreme of the interpretations regarding the nature of mental organization As the name of the theory indicates, intelligence is said to be constituted of a multitude of separate factors, or elements, each one being a minute element of ability Any mental act, according to this theory, involves a number of these minute elements operating together Any other mental act also involves a number of the elements in combination Hence, if performances on these two mental tasks are positively correlated, the degree of correlation is due to the number of *common elements* involved in the two acts If two types of mental activities, A and B, are more highly correlated than are A and C, the reason, according to the multi factor theory, would be that the first pair has more elements in common than does the second pair According to this theory, then, there is really no such factor as "general intelligence", there are only many highly specific acts, the number of such depending upon how refined a classification we might wish to make and are capable of making

Thorndike's is really an "atomistic" theory of mental ability He adds, however, that certain mental activities have so many of their elements in common that it is useful to classify these tasks into separate groups to which special names are given, for example, verbal meaning arithmetical reasoning, comprehension, visual perception of relationships, and others Consequently, in constructing a mental test, it appears even to Thorndike himself that his "atomistic" theory and the multitude of minute elements of ability are of less practical significance than the conception that many of them operate together in any situation demanding intelligence This is illustrated by Thorndike's test designed to measure ability to deal with abstractions His test is composed of four parts sentence completion (C), arithmetical reasoning (A), vocabulary (V), and following directions (D) This instrument is known as the ***CAVD test*** It is not claimed by Thorndike that these four sets of items encompass the entire range of ab-

[12] This of course is true of all sciences

stract intelligence They represent and sample only certain parts, but because of the very significant correlations between all types of meas ures within the tested range, it is held, the other aspects of abstract intelligence can be estimated with satisfactory accuracy from those portions that are actually measured by this test

*The Two-factor Theory.* *Opposed to* Thorndike's theory of the nature of intelligence is Spearman's two factor theory, which stands at the other extreme of interpretations According to Spearman, all in tellectual activity is dependent primarily upon and is an expression of a *general factor* common to all mental activity This factor, designated by the symbol *g* is possessed by all individuals, but in varying degrees of course, since people differ in mental ability, and it (*g*) operates in all mental activity, though in varying amounts, since mental tasks dif fer in respect to their demands upon general intelligence Spearman characterized this general factor as mental energy, because in the realm of intelligent activity, he maintained, it has a role similar to that of physical energy in the physical world Like all other scientific concepts the general factor can be observed and known only through its specific manifestations—in this instance, through psychological tests After analyzing tests with varying amounts of the general factor from high to low, Spearman concluded that the principal distinguish ing characteristic of tests highly 'loaded with *g* is that they require insight into relationships—what he called 'the eduction of relations and correlates' For example, in solving an arithmetical problem the subject has to grasp the relationships between the data presented organize them with reference to the propositions given in the problem, and deduce a correct answer The *g*-content in this task is high By contrast, if the subject merely has to repeat a table of multiplications or add a few numbers—both of which can be learned by rote—no insights are necessary and no relationships need be grasped In this task, the amount of *g* involved is very small [13]

Spearman postulated the *g* factor, in the first place, to explain correlations that he found to exist among diverse sorts of perceiving

[13] When using an individual test like the Stanford Binet or the Bellevue it has often been observed by examiners that a subject who is unable to solve an arithmetical *problem*—unable even to make a start toward a solution—may still be able *to perform the separate arithmetical processes involved In Spear* man s terms such an individual is unable to educe the necessary relations and correlates for lack of the necessary amount of *g*

knowing reasoning, and thinking as illustrated in Table 9 That is to say, he concluded that all mental activity is to some extent dependent upon and an expression of this general factor, and the magnitude of the correlation coefficient found between any two forms of mental activity reveals the extent to which this *g* factor is operative in each and common to both Thus, the amount of the general factor operating in each activity will determine the size of the correlation between the two mental activities being measured The types of materials used in current tests of intelligence—word meaning, arithmetical reasoning, sentence completion, reasoning by analogy, paragraph interpretation,

TABLE 9

*Intercorrelations of Subtests*[14]

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| (1) Analogies | | 50 | 49 | 55 | 49 | 45 | 45 |
| (2) Completion | 50 | | 54 | 47 | 50 | 38 | 34 |
| (3) Understanding paragraphs | 49 | 54 | | 49 | 39 | 44 | 35 |
| (4) Opposites | 55 | 47 | 49 | | 41 | 32 | 35 |
| (5) Instructions | 49 | 50 | 39 | 41 | | 32 | 40 |
| (6) Resemblances | 45 | 38 | 44 | 32 | 32 | | 35 |
| (7) Inferences | 45 | 34 | 35 | 35 | 40 | 35 | |

perception of relationships in geometric forms, picture completion and others—all show significant degrees of positive correlation with one another Spearman and his supporters at first ascribed this fact to the presence of *g*, in greater or lesser amount, in all of them Later researches led them to conclude that certain group factors" are also present in some mental activities These are the factors that occur in more than one type of test item but in less than all of any given set of tests The general factor, however, still remains the primary and pervasive one

Since the intercorrelations are by no means perfect, Spearman postulated the existence of specific factors, called *s* factors, each of which is specific to a particular type of activity Thus, the two factor theory states that all mental activities have in common some of the general factor, each mental activity might also be a member of a 'group', and each has also its own specific factor Of the kinds of factors, the general one is regarded as the essential measure of intelli-

[14] From C Spearman *The Abilities of Man* New York Macmillan 1927 p 149 (By permission )

gence, accordingly a sound test of intelligence is one that will sample adequately the g factor in a variety of activities, and the best test materials are those which call for the largest *amount of the general factor* And the largest amounts of the general factor are believed to be demanded by those types of test materials that have the higher intercorrelations with one another

As a matter of fact, since the beginning of modern mental testing, psychologists have proceeded, at least implicitly, on the assumption that all forms of mental activity have something in common—that they are similar in certain basic respects Otherwise, psychologists could not have justified their practice of testing together in a single instrument such diverse mental activities as defining words, solving arithmetical problems, finding similarities and differences, repeating digits forward and backward, *completing sentences in a meaningful* manner, perceiving geometric forms, etc All of these, and the others used, must have been regarded as being measures, to a greater or lesser degree, of general intelligence From the total performance on these tests, it was believed that an individual s level of general intelligence would emerge *Therefore, psychologists believed they were justified in adding up the test items correctly passed in the several types of activity and deriving a single total score to represent an individual s general intelligence level* [15] This is the actual practice being followed in nearly all tests, including individual as well as group scales of mental ability [16]

The practical implication of the Spearman two-factor theory is clear, so far as test construction is concerned A test conforming to this theory would be one whose materials and several parts are saturated with the general factor so that measurement thereby would cause

[15] While this practice is not being discontinued and should not be increased emphasis is now being placed on *the desirability of representing each individual* by means of a test profile where possible as well as by a general index There are some psychologists however, who would abandon the use of all indexes of general level and would substitute a profile representing the individual s relative rank in each of the specific types of test materials being used e g numerical ability word meaning, spatial perception and the like

[16] In addition to *g* and *s* Spearman and others have found by further analysis of experimental results that there are some nonintellectual factors—such as volition interest persistence—which influence a person s effectiveness Spearman and adherents of his theory have also discerned a few *groups* of factors that are intermediate between *g* and the highly specific *s* They suggest that musical aptitude and mechanical aptitude are of this type

the testee's level and quality of *g* to emerge, while the effects of specific factors (*s*) would be canceled out Thus, the net result of the test would be a measure of *g* To achieve this would require a skillful selection and development of test problems and parts that are significantly intercorrelated, which at the same time satisfy the practical criteria of intelligent activity Such a test, presumably, would yield an index which reflects the caliber of a particular mentality working as a whole

**The Group-factor Theory.** Intermediate between the theories of Thorndike and Spearman are the group factor theories, from among which we select for presentation that of Thurstone because it has been most highly developed, has received most consideration and has resulted in the construction of a set of measures called *tests of primary mental abilities*

According to the group-factor theory, intelligent activity is not an expression of innumerable highly specific factors, as Thorndike claimed Nor is it the expression primarily of a general factor which pervades all mental activity and is the essence of intelligence, as Spearman held Instead, the analyses and interpretations of Thurstone and others led them to the conclusion that *certain* mental operations have in common a 'primary' factor which gives them psychological and functional unity and which differentiates them from other mental operations These mental operations, then, constitute a "group" A second group of mental operations has *its* own unifying "primary" factor, a third group has a third, and so on In other words, there are a number of groups of mental abilities (the number being as yet undetermined) each of which has its own "primary' factor, giving the group a functional unity and cohesiveness Each of these 'primary" factors is said to be relatively independent of the others

After administering a large variety of types of test materials to college students and to high school and eighth grade pupils, and after making correlational analyses of the results, Thurstone and his collaborators concluded that six 'primary factors emerged clearly enough for identification and use in test design and construction They are, briefly, the following [17]

[17] From L L Thurstone and T G Thurstone *The Chicago Tests of Primary Mental Abilities* Manual of Instructions Chicago Science Research Associates, 1943 p 7 See also L. L Thurstone *Primary Mental Abilities* Psychometric Monograph No 1, Chicago University of Chicago Press 1938 L. L. Thurstone

The Number factor (N) "ability to do numerical calculations rapidly and accurately"

The Verbal factor (V) "found in tests involving verbal comprehension"

The Space factor (S) "involved in any tasks in which the subject manipulates an object imaginally in space"

The Word Fluency factor (W) involved whenever the subject is asked to think of isolated words at a rapid rate"

The Reasoning factor (R) "found in tasks that require the subject to discover a rule or principle involved in Series or groups of letters" Although it is believed both induction and deduction are involved it seems that induction is the more significant here

The Rote Memory factor (M) involving the ability to memorize quickly"

In spite of the fact that "primary" mental abilities (or factors) were originally said to be functionally independent of each other, it was actually found that they are positively and significantly intercor-

TABLE 10

*Intercorrelations of Subtests*[13]

| | *N* | *W* | *V* | *S* | *M* | *R* |
|---|---|---|---|---|---|---|
| N | | | | | | |
| W | 41 | | | | | |
| V | 40 | 54 | | | | |
| S | 28 | 17 | 16 | .. | | |
| M | 31 | 36 | 35 | 13 | . | |
| R | 53 | 49 | 59 | 29 | 39 | .. |

(N, number facility, W, word fluency V, verbal meaning; S spatial perception, M, rote memory; R, reasoning)

related, as shown in Table 10 This must mean that the "primary" and presumably independent factors are not the only factors at work in the mental activites required by the tests There must be some other

and T G Thurstone, *Factorial Studies of Intelligence* Psychometric Monograph No 2, Chicago University of Chicago Press 1941 Some modifications of factors have been introduced in the most recent issues of these tests for younger subjects The six named above are still regarded as established factors.

The "primary" mental abilities do not include the entire range of human abilities They do not include mechanical musical or artistic aptitudes The "primary" abilities involve largely though not entirely those which are needed in "abstract intelligence" and in the academic types of learning.

[13] From L. L. Thurstone and T G Thurstone *The Chicago Tests of Primary Mental Abilities* Manual of Instructions, Chicago Science Research Associates, 1943 (By permission )

lents for the scores on the scale for ages 11 to 17, while for the younger age levels both MA units and quotients are provided

As is so often the case in scientific problems—especially in the relatively new ones—divergent theories in time tend to come into closer agreement The Spearman Two Factor Theory now recognizes that some group factors should be posited to explain test findings, but em phasis is upon the g factor Perhaps the Spearman theory may now be renamed ' The General Factor Group Factor Theory," while the other might be renamed ' The Group Factor General Factor Theory The narrowing of differences between the two theories represents sig nificant scientific progress

## FACTOR ANALYSIS

The two factor and the group factor theories are the two most prominent examples of doctrines emerging from the methods of factor analysis Although this subject is highly technical, it is desirable to explain it somewhat more fully at this stage

The technique is essentially a search for the psychological functions which are at the basis of and determine test performance All tech niques of factor analysis are statistical and based upon the correlation coefficient After the statistical calculations have been made, it is nec essary for the investigator to bring to bear his psychological insights to interpret and name his statistical findings Tests contain a variety of items What psychological functions do the various types of items have in common? Are there functions in common between various tests of verbal performance? Between verbal and numerical? Between spatial perception and numerical ability? Between reasoning with verbal and with nonverbal materials? These are among the questions the factor analyst seeks to answer After he has found his answer at least tentatively, he proceeds to construct a scale in which items are included and so grouped as to measure only, or almost solely, the factors he has segregated from his preliminary testing and statistical analysis

The factor analyst does not begin with a definite set of preconceived mental functions He tries to discover which psychological functions, or components, are necessary to explain his data Yet, it should be noted he must at the very outset have some conception of the kinds of test items to include in preliminary experimentation Thus what he ultimately distills out as factors is basically dependent upon his origi-

nal conceptions regarding his preliminary items The factor analyst, in seeking the components of "intelligence," for example, does not start with tests of color perception, tone discrimination, or finger dexterity

**Two-factor Theory.** We have already stated Spearman's *two-factor theory* It will be helpful now to describe in more detail the reasoning whereby the theory was arrived at Spearman, in his early experimentation, was impressed by the fact that all the intercorrelations were positive in a table of coefficients for various types of items He was also impressed by what appeared to be a hierarchy of coefficients in the rows and columns of the table, not perfect gradations, but strong evidence of proportional gradations He therefore offered a hypothetically perfect table of correlation coefficients to illustrate his point (Table 11)

TABLE 11

Spearman s *Hypothetical Table of Correlations*[4]

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| (1) Opposites | | 80 | 60 | 30 | 30 |
| (2) Completion | 80 | | 48 | 24 | 24 |
| (3) Memory | 60 | 48 | | 18 | 18 |
| (4) Discrimination | 30 | 24 | 18 | | 09 |
| (5) Cancellation | 30 | 24 | 18 | 09 | |

Not only are the coefficients positive and in a decreasing order along rows and columns, but theoretically any two columns (or rows, since the table is symmetrical about the diagonal which contains the self correlations) are in direct proportion The criterion of proportionality requires that the following correlational relationships should hold

$$\frac{r_{13}}{r_{23}} = \frac{r_{14}}{r_{24}} = \frac{r_{15}}{r_{25}}$$

Taking only the first two ratios,

$$\frac{r_{13}}{r_{23}} = \frac{r_{14}}{r_{24}},$$

and multiplying by the denominators, we have

$$r_{13}\, r_{24} = r_{23}\, r_{14}$$

[4] From C Spearman, *The Abilities of Man* New York Macmillan, 1927, p 74 (By permission )

Transposing, we get

$$r_{13}\,r_{24} - r_{23}\,r_{14} = 0$$

From this *tetrad equation* (so called because the test correlations are dealt with in sets of four) may be obtained what is known as the *tetrad difference*

The tetrad equation may be written for the combination of any four tests By rearranging the four coefficients, three tetrad differences may be obtained for every combination of four tests Thus, using $t$ as the notation for tetrad difference

$$t_{1234} = r_{12}\,r_{34} - r_{13}\,r_{24}$$
$$t_{1243} = r_{12}\,r_{34} - r_{14}\,r_{23}$$
$$t_{1342} = r_{13}\,r_{24} - r_{14}\,r_{23}$$

Theoretically, the tetrad difference criterion is satisfied if $t$ is zero When it is zero, Spearman and others have offered mathematical evidence to demonstrate that a single common factor can account for the relationships among the four tests, or variables But in actual fact, the difference is rarely if ever zero However, if the differences are close to zero, we may also conclude the criterion is satisfied, since correlations between tests will be affected by errors of measurement due to chance and accidental factors[25] (Cf discussion of reliability in Chapter 1 ) Furthermore, the correlation coefficients would also be affected by the operations of the *specific* factor in each test The specific factor was postulated to explain, in part at least, the tetrad differences that were greater than zero

If more than four tests are being examined to disclose the functions involved, we may substitute, say, tests numbers 5 and 6 in the tetrad equations in place of numbers 3 and 4 Thus we would be analyzing tests 1, 2, 5, 6 Then if the tetrad difference criterion is satisfied, it would be concluded that the functions common to 1 and 2 are identical with those common to 5 and 6 Assuming that the tetrad difference criterion was satisfied also for tests 1, 2, 3, 4, then the same functions (or factor) are said to be common to the six variables The same reasoning may be applied to any number of tests

The principal conclusion drawn by Spearman and some others, after their analyses, was that the degree of correlation between any

[25] Formulas are provided for calculating probable errors of tetrad differences comparison of a tetrad difference with its probable error (PE) enables one to decide whether it differs significantly from zero

two tests is dependent upon the extent to which *g* is involved in each Subsequent investigations showed, however, that *some* test intercorrelations may include their own common factors beyond the single common *g* It was necessary, therefore, to postulate the operations of *group factors* each group being effective in two or more tests, but not in all of them Spearman and others recognized such group factors as numerical verbal, speed, mechanical, imagination, and attention In addition Spearman had postulated three nonintellective factors which influence one s mental effectiveness perseveration (*p*), oscillation (*o*), being one s variability in performance in continuous mental activity and will (*w*), being one s persistence in effort

The tetrad difference criterion does not in itself show the relative weight or importance of the common factor in each kind of test Following the work of Spearman, therefore methods have been developed for finding the weights (commonly called loadings ) of each factor—general or group—in each of the intercorrelated tests These methods are known as factor pattern analysis [26]

The two factor theory can account for the universal positive correlation coefficients among the various kinds of test items included in scales to measure mental ability, since every form of test requires the operation of *g* to some degree Pooling a variety of kinds of tests in a scale is sound practice, according to this theory, because we thereby approximate a measure of pure *g* Since the *s* factors are uncorrelated within any individual—that is they may be possessed in varying and random degrees by him—they will be a negligible factor in the total performance in a pooled test of general ability, because the varied *s* factors will tend to cancel out one another

Sampling Theory. The two-factor theory has been criticized by some statistical psychologists, notably G H Thomson and L L Thurstone Thomson offers a *sampling theory* [27] to explain the same tables of intercorrelations Briefly, his view is that the coefficients of correlation are the results of common samplings and combinations of independent factors The number of common independent factors utilized by two

[26] For example T L Kelley *Essential Traits of Mental Life* Cambridge Harvard University Press 1935 L L Thurstone *Vectors of the Mind Multiple Factor Analysis for the Isolation of Primary Traits* Chicago University of Chicago Press 1935

[27] G H Thomson *The Factorial Analysis of Human Ability* Boston Houghton Mifflin, 1939, Chapter 3

tests will determine the coefficient of correlation between these two This theory is, of course, the same as Thorndike's, except that Thomson concedes the practical usefulness of a concept like *g* Thomson also adds that if several tests call upon many elementary factors in common, they will not only have a very marked or high coefficient of correlation, but they will give the appearance of having *one* common comprehensive factor Also, Thomson's theory maintains that if several tests draw upon a relatively smaller number of the elementary factors in common, we have then group factors—that is, a limited number of factors that enter into performance on types of tests which are distinguished by the fact that they have certain mental processes in common but do not share a very large number of elementary factors or a universal *g*

While both theories require that a scale to measure general mental ability should pool a variety of types of tests differing in content and mental processes employed, the two factor theory would seek subtests (parts of the scale) that have high intercorrelations, whereas the sampling theory would seek subtests having low intercorrelations among themselves but high correlations with the criteria of validity

**Group-factor Theory.** As already stated, Thurstone and others believe that a *group factor theory* fits the facts best and is most useful in testing practice Their view differs from Thomson's in that they reject the theory of a very large number of independent factors As previously explained, a group-factor is conceived of as an operational concept to account for correlations of performance within only a limited group of tests [28] Several different groups of factors are necessary to account for all mental activity, plus the more recently added second-order *g* factor, which Thurstone states may be more "central" in character and more "universal" in influence

Thurstone has contributed very significantly to the methodology of group-factor analysis, particularly his geometric methods, and from his analyses, as we shall see, has emerged a scale to test mental ability *This volume is not the place to present his techniques,* we shall merely state his purposes [29]

[28] Most recently, some group-factor theorists have characterized "primary factors" as facilities of the mind and as media of expression.

[29] A number of others have made significant contributions to factor theory, especially K. J Holzinger H Hotelling R C Tryon H E Garrett C L Burt, J C Flanagan J P Guilford, and P Vernon

Three objectives, according to Thurstone, are to be achieved by factor analysis (1) determination of the smallest number of primary mental abilities to be postulated as an explanation of tables of intercorrelations, (2) determination of the amount of each primary ability that is involved in each test, and (3) determination of regression

TABLE 12

*The Two Factor Pattern*

| *Test* | *General Factor* | *Specific Factors* |
|---|---|---|
| 1 | x | $S_1$ |
| 2 | x | $S_2$ |
| 3 | x | $S_3$ |
| 4 | x | $S_4$ |
| 5 | x | $S_5$ |
| 6 | x | $S_6$ |

*The Group-Factor Pattern*

| *Test* | *Group Factors* | | | |
|---|---|---|---|---|
| | *A* | *B* | *C* | *D* |
| 1 | x | | x | x |
| 2 | x | x | x | |
| 3 | | x | | x |
| 4 | x | x | x | |
| 5 | x | x | | x |
| 6 | | x | x | |

*Factor Theories Combined*

| *Test* | *General Factor* | *Group Factors* | | | *Specific Factors* |
|---|---|---|---|---|---|
| | | *A* | *B* | *C* | |
| 1 | x | x | | | $S_1$ |
| 2 | x | x | | | $S_2$ |
| 3 | x | | x | | $S_3$ |
| 4 | x | | x | | $S_4$ |
| 5 | x | | | x | $S_5$ |
| 6 | x | | | x | $S_6$ |

equations whereby the amount of a primary mental ability in an individual can be estimated from tests that draw upon that ability As an illustration, we may consider several tests in which only two group factors are involved An individual might make a high score in these tests either by having a moderately high level of ability in each of Factor I and Factor II, or by having very much of one and little of the other Also, if Factor I carries much heavier weight in the tests

than does II, then a high level of ability on I is more important for high performance on these tests than is a high level of II Thus, the Thurstone method would find the relatively few primary or basic mental abilities, devise a scale to measure all of them, and so organize and score the subtests as to reveal each individual's relative strength in each factor

**Summary.** Methods of factor analysis differ somewhat in their assumptions, and analysts differ somewhat in their interpretations or results, but the general conclusions derived by the several methods of analysis and interpretation do not differ radically All factorial theories now postulate the presence of group factors, although the groups are not always identical and differ in relative emphasis placed upon them by different theories Most theories also find a general factor necessary to explain intercorrelations, although here again emphasis upon the general factor varies All agree that an individual's mental activities are attributable to the various ways in which the general and group factors combine in the performance of varied mental tasks While several methods of factorial analysis are possible, basically the choice between them and interpretations derived through them must rest upon psychological theory and concepts rather than upon statistical methods Factors should not be regarded as fixed, predetermined mental entities The factors that are found are influenced by the ages of the persons tested, by interests, by experience and training, and by the test items originally employed in the preliminary investigations Factorial analysis is a statistical method which provides the means of improving test construction and of classifying test performance

## ILLUSTRATIONS OF FACTORS

The following illustrations will assist the student to grasp more fully the problem of factors in psychological testing As already stated, factor analysis techniques depend basically upon test intercorrelations Tables 13A and 13B show the intercorrelations between two sets of four subtests of the Wechsler scale for children

Since all the coefficients in Table 13A are positive and quite marked, we conclude that all four tests have much in common, but that Vocabulary, Information, and Similarities have somewhat more in common with one another than they do with Comprehension But

since all the coefficients are far from perfect (+1 00), we are warranted in using all four to sample the testee s mental abilities, rather than only one or two to the exclusion of the others The fact that these four tests are not perfectly correlated—nor nearly so—might be due to one of these possibilities (1) that each test samples the *g* factor in different amounts, plus its own specific factor, or (2) that the tests have *g* in common but each test samples also one or more *group factors,* though not necessarily the same ones, or (3) that each has many highly specific factors in common with every other one, as

TABLE 13A

*Intercorrelations of Four Subtests of the Wechsler Intelligence Scale for Children* [30]

| | *Vocab* | *Info* | *Sims* | *Comp* |
|---|---|---|---|---|
| Vocabulary | – | 74 | 66 | 60 |
| Information | | – | 67 | 61 |
| Similarities | | | – | 61 |
| Comprehension | | | | – |

TABLE 13B

| | *Obj Assemb* | *Comp* | *Arith* | *Digit Sp* |
|---|---|---|---|---|
| Object Assembly | – | 13 | 20 | 13 |
| Comprehension | | – | 46 | 28 |
| Arithmetic | | | – | 40 |
| Digit Span | | | | – |

well as unique factors A technical factorial analysis would go beyond this inspection analysis in an effort to determine which of these three hypotheses is the most plausible one, and to determine to what extent performance on each of the four tests calls upon whatever factors —*g* or others—might be inferred from the statistical analysis

Table 13B, by contrast, shows four subtests that have low intercorrelations Of the six coefficients, only two ( 40 and 46) are large enough to suggest that the subtests involved have much in common as regards psychological functioning The coefficient of 46 between Comprehension and Arithmetic is attributable to the demands that both of these tests make upon reasoning ability, or, more specifically, ability to analyze a set of given material and then reorganize the elements toward the solution of the specified problem The coefficient

[30] From the *Manual* Psychological Corporation.

of .40 between Arithmetic and Digit Span is attributable, it appears from the characteristics of the two tests, to facility with numbers and ability in immediate recall (as contrasted with delayed recall) The remaining four coefficients are so low as to suggest that the tests concerned have little dependence upon common functions (whether *g* or other factors), and that each makes demands upon some factor or factors not called upon by the others Here again, a factorial analysis would attempt to identify more precisely the factors involved, but in so doing the analyst would have to apply his knowledge of psychological functioning to the items that are shown by the analysis to cluster together

Figure 3 1 shows in graphic form how two and three tests *might be* interrelated As the number of types of tests increases, the possible factor interrelationships may become more numerous and complex, though it is extremely improbable that *no* overlapping whatever of factors would be found in measuring human abilities by means of two or more different types of tests In view of the more recent partial reconciliation of the group-factor and the general factor theories, the illustrated overlappings are most probably attributable to *g*

The possible factor interrelationships of the parts of Figure 3 1 are

A Each test is factorially independent of the other The factor or factors in each are unique to it, either as "group" factors or as "specific" factors

B The overlapping shaded area indicates a factor or factors common to both tests This may be *g* or a group factor The unshaded area indicates factors unique to each, either specific or group, or both When a number of diverse tests show some overlapping among all of them, the soundest inference is that a *g* factor accounts for the common ground

C. In this instance the tests may be measuring only the general factor, or *g* plus the same group factor, or just the same group factor There are no unique factors It is extremely improbable, in this situation, that the general factor is not involved If numerous pairings of different tests showed this relationship, the soundest inference would be that a general factor is being measured.

D Each of the three tests is factorially independent of the other two The uniqueness of each may be due to group or specific factors, or to both

E The overlapping of 1 and 2 here may be attributable to *g* or to a group factor Test 3 is independent of the others The nonoverlap-

ping segments of 1 and 2 may represent separate group factors or specific factors in each test

F In this figure overlapping between 1 and 2 and between 2 and 3 *is attributed to one or more group factors* but different ones in each case since there is no common ground between all three tests

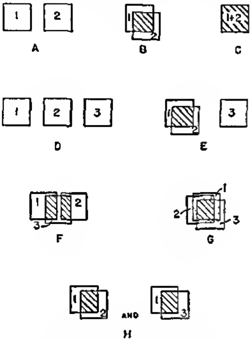



FIG 3 1 *Possible Intercorrelations Between Two and Three Tests*

The unshaded areas would represent special factors or group factors, or both which are not shared with either of the other two tests

G Here there is some common ground in all three tests (shaded area) which is interpreted as showing the presence of the *g* factor The dotted areas show group factors shared by only two of the tests The unshaded areas represent either specific factors or unique group factors or both

H This figure represents three tests that have only the general factor in common Both Tests 2 and 3 have the same amount and the identical area in common with Test 1 hence they have the same amount and identical area in common with each other

These graphic illustrations of correlations and factor "loadings" derived from statistical analysis serve three purposes (1) They demonstrate the complexity of the problem of determining interrelation-

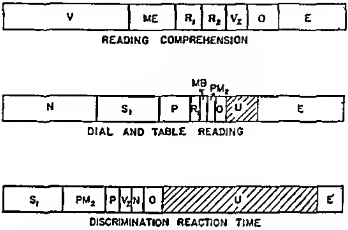



FIG 3 2 *Diagrams of the Component Variances of Three Army Air Force Classification Tests* (From Guilford, op cit p 86) *The letters stand for*

- $V$ *verbal-comprehension factor*
- $ME$ *mechanical-experience factor*
- $R_1$ *reasoning I (general reasoning) factor*
- $R_2$ *reasoning II (common to analogies tests) factor*
- $V_2$ *visualization factor*
- $O$ *other common factors, each with variance too small to mention separately*
- $U$ *unknown common factor or specific factor variances*
- $E$ *error variances*
- $N$ *numerical factor*
- $S_1$ *space I (spatial relations) factor*
- $P$ *perceptual speed factor*
- $MB$ *mathematical background factor*
- $M_2$ *memory II (visual memory) factor*
- $PM_2$ *psychomotor II (precision) factor*

interpretations help to make clear the reasons why the most valid and useful tests within a given category (e.g., intelligence) have much in common as regards psychological functioning and as regards test items themselves.

Finally, Figures 3.2 and 3.3 illustrate elaborate factorial analyses of tests which have been statistically fractionated.[31] These indicate the probable quantitative portions of each of the several factors in eacb of the tests. Such analyses do, undoubledly, provide insights into the

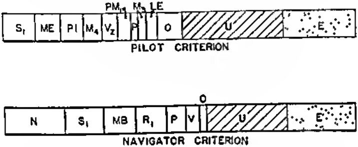



FIG. 3.3. *Diagrams of the Component Variances of Pilot and Navigator Training Criteria.* (From Guilford, op. cit., p. 86.) *Letter symbols as defined with Figure 3.2, except for some additional ones:*

*PI* pilot-interest *factor*
*$M_4$* memory *IV* (content-memory) *factor*
*$M_3$* memory *III* (picture symbol association) *factor*
*$PM_1$* psychomotor *I* (coordination) *factor*
*LE* length-estimation *factor*

psychological operations that combine in performance on the tests. Test construction is thereby facilitated. It should not be assumed, however, that each of these factors exists or operates independently. We may look at the factors in the Reading Comprehension test as an example. We note that "verbal comprehension" is the largest single factor; then we have, in order, "mechanical experience," "reasoning I" and "reasoning II." It is very doubtful that these four factors can or should be separated functionally. Mechanical experience will have been significant in the comprehension of the verbal materials of this test and will have contributed to the verbal competence of the testee

[31] J. P. Guilford, "Factorial Analysis in a Test-Development Program," *Psychological Review*, Vol. 55, 1948, pp. 79–94.

in the particular area covered by the test Reasoning, of whatever kind, is basically a matter of problem-solving ability, whether with the use of concrete objects or with abstractions (words and numbers) In this test of Reading Comprehension, the examinee's ability to reason with the problems presented will be dependent in part upon his mechanical experience (contributing to his comprehension) and to his knowledge of the terms used in the problems Conversely, the extent and quality of the vocabulary he acquires and the degree to which he benefits from his mechanical experience will depend in part upon his present and potential reasoning capacity

## IMPLICATIONS

The hypotheses as to the nature of mental abilities have been arrived at by means of several methods of statistical analysis and through partially different interpretations placed upon much the same data by different investigators Regardless of which of the hypotheses an author of a test follows, the instrument he develops will have very much in common with those constructed by authors who base their tests on one of the other hypotheses In many respects, the processes of standardization will be the same, the same basic principles of constructing and testing will have to be observed A variety of mental activities will have to be sampled, in the case of the multi-factor theory, in order to sample an adequate and representative number of the many minute factors, in the case of the group-factor theory, in order to sample the "primary" abilities and those "second-order" factors that might be found subsequently, in the case of the two factor theory, in order to get an adequate sampling of the general factor

The main practical differences arising from the theoretical differences will be found in the tests based on the group-factor theory, as compared with others The differences will be these the parts of the test based on group-factor theory must correspond with the factors or "primaries" and they must try to measure these factors in as 'pure" a form as possible, the subtests in a scale based upon group factory theory should have low intercorrelations, the test based on group-factor theory will emphasize the separate scores on each of the ' primaries" and will provide a "mental profile," even though the group-factor test might also provide an over all index The Binet type of test, and other *g* factor tests, on the other hand, consists of a variety of test materials which are a composite of abilities, yielding a mental

age and an intelligence quotient Most group tests [32] while arranging their items according to type (sentence completion arithmetical reasoning word meaning picture completion form perception, etc ), are not organized on the basis of specifically defined factors, and, like the Binet type, they generally yield a single index of relative rank

Since the several hypotheses regarding the nature of intelligence have thus far produced relatively few differences in practical test construction and application, the reader might well ask Why, then, be so concerned with definitions and theories, when tne end results are oot radically different? The answer to this question has several aspects First, the student should be familiar with the thinking of psychologists in this field, as a background for his better understanding of the tests themselves Second it is through the interaction of the theoretical and applied that improvemeots and advances will be made Third it is possible that one or more of these theories will have increasing influence, in the future upon test construction, testing practice and test interpretation [33]

With these definitions and theories of intelligence in mind, we turn now to an examination and evaluation of some of the representative current tests of intelligence

---

[32] *Excepting the omnibus type in which items of various mental operations* are placed in regular or irregular order instead of being grouped in subtests each containing items of a single kind

[33] A useful presentation of some problems of test construction will be found in J Loevinger *A Systematic Approach to the Construction and Evaluation of Tests of Ability* Psychological Monographs Vol 61 No 4 1947 See also L L Thurstone "Psychological Implications of Factor Analysis" *The American Psychologist* Vol 3 1948 pp 402–408

# 4.

## THE BINET SCALES

### HISTORICAL BACKGROUND

The development of mental tests was motivated by psychologists' interests in and attempts to measure individual differences in abilities Experimental work in testing was begun in the 1860 s by Francis Galton the English geneticist and eugenist who wished to study the effects of heredity and environment upon the development of mental abilities and thus the possibilities of supplanting inefficient human stock by better strains It was necessary, therefore that he develop methods of testing psychological traits which would be more objective and valid than any means then available

It was to be expected—in view of Galton s background as a biologist, the ascendance of biological science during this period and the emergence of experimental psychology as a separate discipline—that Galton should attempt to measure intelligence by means of the tests of imagery and sensory discrimination which he devised For example he used a test for the measurement of the delicacy of weight discrimination and what is known as the Galton whistle for measuring susceptibility to high tones Galton also suggested devices to be used for testing the other senses In addition he devised tests of imagery which were intended to discover the fidelity and detail with which a person could recall an earlier experience

Apparently Galton assumed that the simpler measurable sensory capacities would show a significant correlation with intelligence and that if these simple sensory measures were obtained they would afford a means of judging or predicting an individual s intellectual ca-

pacity Although it has since heen demonstrated that these measures have very little or no value for evaluation of the higher and more complex processes which we call 'intelligence," Galton s work did, nevertheless, affect the nature of test experimentation until about 1900, when the influence of Alfred Binet, the French psychologist, was felt

The kinds of psychological tests which were being used prior to 1900 are well represented hy those with which J McKeen Cattell was experimenting in the United States[1] Ten of his most highly regarded tests at that time were the following

(1) Strength of grip, using the dynamometer

(2) Rate of movement the quickest time in which the hand can he moved through a distance of fifty centimeters

(3) The smallest perceptible distance between two points on the skin, known as two *point discrimination* '

(4) Amount of pressure necessary to cause pain hy exerting pressure upon the forehead with a strip of hard rubber

(5) The smallest difference in weight which can be discerned, measured by requiring that two weights be lifted in succession

(6) Speed with which an individual can react to a sound

(7) Speed with which an individual can name ten specimens of four different colors arranged in haphazard order

(8) The accuracy with which an individual can bisect a fifty-centimeter line

(9) The accuracy with which an individual can reproduce an interval of ten seconds

(10) Immediate rote memory, using a series of consonants

It is obvious that these are, for the most part, simple sensory and motor tests, with a little rote memory added

Other investigators in the United States and abroad were also experimenting with psychological tests, employing methods and materials similar to those of Galton and Cattell Jastrow gave tests involving hoth touch and vision, tests of vision alone, tests of memory, and tests of reaction time Boas made physical measurements of children and also tested their vision, hearing, and memory At the same time he obtained teachers' estimates of their pupils' intellectual acuteness Gilbert used measures of height, weight, and lung capacity, tests of sensation, rapidity of tapping, reaction time, memory, and

[1] J McK Cattell, "Mental Tests and Measurements" *Mind* Vol 15, 1890, pp 373–81

suggestibility. He also obtained teachers' ratings of the pupils' mentality.

Several other investigators during this period were beginning to introduce somewhat more complex materials and methods. Kraepelin and Oehrn, in Germany, used tests of perception involving the counting of letters, cancellation of letters, and detection of errors on the printed page, tests of memory involving digits and nonsense syllables, tests of association and of motor functions.

Muensterberg also devised tests of a more complex kind: reading aloud rapidly, giving rapidly the colors of named objects, rapidly naming and classifying animals, plants, and minerals, naming rapidly and classifying cloth, food, and parts of the body, naming rapidly ten simple designs and ten squares of colors, tests of addition, rapidly counting angles in irregular polygons, naming different odors. He also employed tests of memory for digits and letters after a single presentation, bisecting, judging, and reproducing lengths of lines, locating a sound, constructing a square and an equilateral triangle, with only the base of each given.

Some of Muensterberg's tests, it will be noted, were more complex and more varied than those of other experimenters, yet, like those of others, his were tests essentially of simple psychological processes, with a premium placed upon speed.

These relatively simple tests of sensory, motor, and memory capacities proved to be of very little value as measures to reveal intelligence. In the first place, their *intercorrelations* were very low, ranging generally from zero to only .20. And, in the second place, the results of these tests, when correlated with academic performance, yielded correlation coefficients of much the same magnitude, many of them being less than .10—hence useless for purposes of prediction. As a matter of fact, experimentation in the years that followed the early investigations has consistently confirmed the negligible or very low correlations found to exist between sensory and motor capacities, on the one hand, and the higher more complex functions, called "intelligence," on the other.

It is now generally recognized by psychologists that intelligence has little relationship to the elementary sensory and motor processes, and but a very moderate relationship indeed to capacity for rote memory (a correlation of about .30). Many infra human animals have keen sensory discrimination. Mentally deficient children in the higher levels

of defect and children in the "borderline" group are not very inferior to normal children in respect to skin sensitivity, visual acuity, auditory acuity, reaction time, etc Nor are intellectually gifted children superior to average in these respects But in the capacities to learn, to organize and direct thinking, to adapt behavior, to comprehend problems and deal with abstractions, in levels of information possessed, in extent of curiosity about one's environment, these groups do differ very markedly indeed

The reader should bear in mind these early kinds of tests, not only for historical purposes, but also in order to compare the early efforts with currently available tests and to be more clearly aware of the direction in which psychological testing has been moving

## THE EARLY WORK OF ALFRED BINET

The earlier work of Binet, the French psychologist, was along much the same lines as that of the American and German psychologists already mentioned He used tests of tactual discrimination, reaction time, visual discrimination, auditory discrimination of time intervals, reproducing letters and numbers from memory, and so on But though he experimented with these materials until about 1900, he had begun to doubt, some years earlier, the value of continuing with them

Although some of the mental activities which Binet proposed to measure and with which he was experimenting were as yet vague, he did nevertheless point out the direction in which mental tests should and in fact did develop, that is, that the higher and more complex mental functions must be measured rather than the simpler sensory and motor processes

Binet and his collaborators objected to the kinds of psychological tests which followed Galton's work, on the ground that they were too simple in character and would contribute little to the understanding of differences among persons in respect to the higher mental functions Binet maintained, furthermore, that intelligence is expressed *not* in the form of simple, segmental responses, but rather as a combined mental operation wherein whatever processes are involved operate as a *unified whole*

It is in these higher functions that individual differences are most marked, it is these which distinguish individuals most significantly and characteristically in daily activity, whereas it is in the simpler

sensory and motor processes that persons differ least significantly Binet was quite ready to grant that the simpler processes, such as those already mentioned, lent themselves to more precise measurement and, therefore, yielded more constant results Yet his interests were strongest in individuals rather than in sensory and motor processes as such Thus, he was ready to sacrifice the greater quantitative precision of sensory motor tests in order to obtain more valid evidence of the integrated mentality of the individual, for he argued that in the measurement of the higher functions, the greatest precision of measurement, though desirable, was not as essential as in measuring the simpler processes, because of the very fact that individuals differ much more markedly in the former

FIG 4 1 *Alfred Binet (1857–1911)*

Binet emphasized that his proposed scale, testing the higher and more complex mental functions, would not measure in a physical sense, in the same way, for example, that the length of a line is measured His tests would, however, yield "a classification, a hierarchy among diverse intelligences, and for the necessities of practice this classification is equivalent to a measure " [2]

Binet and his collaborators were interested in establishing the ex-

[2] A Binet *The Development of Intelligence in Children* trans. by E. S Kite, Vineland Training School, 1916, p 40

tent and nature of variations of mental functions from one individual to another and in the determination of the interrelations of the several functions *within* the individual In 1896 therefore, Binet and Henri (a collaborator) published their studies of the following functions memory, the nature of mental images imagination attention, comprehension suggestibility esthetic appreciation moral sentiments, muscular strength strength of will motor skill and visual judgment These they believed were functions which differ much from one individual to another and are such that knowledge of their state for an individual gives us a general idea of this person permits us to distinguish him from other individuals belonging to the same milieu [3] Here then we have the beginnings of the tests which a few years later proved so useful in the construction of the Binet scales and subsequently in the several revisions developed for use in the United States

In 1904 a practical situation arose in which Binet had an opportunity to apply his principles with regard to differentiating individuals and to make an outstanding contribution to the study of mental abilities and individual differences The French Minister of Public Instruction appointed a commission to recommend measures to be taken in the education of mentally subnormal children in the schools of Paris for it was recognized that these children were unable to profit from the usual instruction The plan was therefore to eliminate subnormal children from ordinary schools and to provide adapted instruction in a special school Admission to the special school was to be based upon a medical and a pedagogical examination Obviously the first device needed was some objective means of selecting pupils of subnormal mentality Subjective opinions were worse than useless for not only was there absence of agreement among different so called experts but serious injustices could result in numerous cases

## THE 1905 BINET SIMON SCALE

It was to meet this problem that Binet constructed his first scale to evaluate childrens intelligence levels This instrument is known as the 1905 Binet Simon Scale [4] In it we find a fundamental

[3] For details see A. Binet and V Henri "La Psychologie Individuelle" *L Annee Psychologique* Vol 2 1896 pp 411–465 For most of Binet s contributions see *L Année Psychologique* Vols 1–17

[4] Called a "scale" because the test items are arranged in order of increasing difficulty

conception underlying all tests by means of which mental abilities of children are measured *The principle is that we may identify differences in mentality, differences in degrees of brightness or dullness, with differences in levels of development as represented by the average capacities of children of various ages* Thus, if we know the levels of intellectual performance of typical, or normal, children at each age, we can determine in the case of any individual child the extent to which his mental development is accelerated or retarded, or whether it is just about at the average level for his age at any given time

In the 1905 scale, this conception was only crudely implemented, but in time it became more precise and has since taken the form of indexes already discussed in Chapter 2 namely, mental age, percentile ranks, decile ranks, standard scores, intelligence quotients, and others

The thirty items, in order of increasing difficulty, comprising the 1905 scale follow [5]

(1) Visual coordination—degree of coordination of movement of head and eyes as a lighted match is slowly moved before subject s eyes

(2) Prehension provoked by tactual stimulus—a small cube of wood is placed on back or palm of the subject s hand to see if he grasps it and carries it to his mouth, and coordination of movements is to be noted

(3) Prehension provoked visually—cube of wood is placed within subject s reach by examiner who notes whether subject grasps it

(4) Recognition of food—a small piece of chocolate and a piece of wood of same dimensions are shown successively, and Signs of recognition of food and efforts to take it are noted

(5) Seeking food when slight mechanical difficulty is interposed —a small piece of chocolate, wrapped in a piece of paper, is given to the subject, and his manner of separating the food from the paper is noted

(6) Execution of simple directions and imitations of simple gestures

(7) Verbal knowledge of objects—parts of the body (head, ear, nose etc ) are indicated by the subject, and common objects (key, string cup) are handed to examiner on request

(8) Verbal knowledge of objects in a picture, as shown by pointing out objects the names of which are given

[5] For a detailed description, see G M Whipple, *Manual of Mental and Physical Tests* Baltimore Warwick and York, 1910 pp 475 ff.

(9) Naming of objects designated in a picture

(10) Comparison of lengths of two straight lines, pointing out the longer

(11) Repeating three digits immediately after hearing the series once

(12) Comparison of weights, identical appearing blocks of wood weighing 3 and 12 grams, 6 and 15 grams, 3 and 15 grams

(13) Suggestibility—asking the subject for an object that is not present (modification of 7), asking subject to point to a nonexistent object in a picture, designated by a nonsensical word (modification of 8) comparison of lines of equal length (modification of 10)

(14) Definitions of familiar objects such as house horse, fork

(15) Repetition of sentences having fifteen words each, after hearing each one only once

(16) Giving the differences between two common objects e g, wood and glass, a fly and a butterfly

(17) Immediate recall of pictures of familiar objects—pictures of thirteen common objects are shown for thirty seconds, after which the subject names as many as he can recall

(18) Drawing from memory two different geometric designs which have been shown simultaneously for ten seconds

(19) Repetition of series of digits, beginning with a series of three and proceeding until the subject's limit is reached

(20) Giving resemblance between common objects e g, a wild poppy and blood, an ant, a fly, a butterfly, and a flea

(21) Rapid comparison of lengths of lines a line of 30 cm is compared with fifteen others varying from 31 to 35 cm, then a line of 100 cm is compared with twelve others varying from 101 to 103 cm

(22) Discriminating and arranging in order five weights—3, 6, 9, 12, 15 grams—all being of equal size

(23) Recall of weights—one of the weights in test 22 is removed, the remaining weights are scrambled, and the subject is asked to identify the missing weight or gap in the series

(24) Giving rhymes to selected words

(25) Sentence completion—supplying the correct word to complete a sentence

(26) Devising a sentence to include three given words e g, Paris, gutter, fortune

(27) Comprehending and giving replies to twenty five problem-questions graded in difficulty e g, What is the thing to do when you are sleepy? Why is it better to continue with perseverance what one has started than to abandon it and start something else?

(28) Reversing the hands of a clock, to be done from memory, e g, giving the time it would be if the large and the small hands were interchanged at four minutes to three The subjects who succeed were given the more difficult problem of explaining why the precise transposition indicated is impossible

(29) Drawing lines to show the folds and cut out of a piece of paper that has been quarto folded and from which a triangular piece has been cut

(30) Giving definitions and distinctions between paired abstract terms e g , sad and bored

Although this set of tests was not separated into age groups, Binet did indicate several differentiating levels Question number 6 was the upper limit of idiots (adult), question 9 was the upper limit of ordinary three-year old children, number 14 was the limit of ordinary five-year-old children, number 16, that of imbeciles (adult), test 23, the most probable limit of morons (adult), although test 27 was regarded as having great value in revealing the moron In addition, the authors reported a number of qualitative and quantitative differences in replies to many of the questions, thus distinguishing between 7 and 9 year levels, on the one hand, and 9- and 11-year levels, on the other

The order of tests in the 1905 scale was experimentally determined, *for it was established after being* used with *children* in the primary schools and in an institution for the mentally deficient (the Salpetriere) The children in primary school were regarded as "normal' on the basis of the fact that they were in grades just right for their ages—neither advanced nor retarded Binet and Simon report that many such children were tested, but norms for the scale were based upon records of only ten cases in each of the following age-groups 3, 5, 7, 9, and 11 years

While admittedly rather crude and tentative, this first scale enabled Binet and Simon to classify idiots, imbeciles, and morons in a more objective manner than had been possible before

Furthermore, in the foregoing list of thirty items, the reader will find many types which have since been developed, standardized, and included in a large number of current psychological tests, from those designed for babies to those intended for adult levels

It is significant to note, also, that while Binet wanted to devise a scale that would yield age ratings, he was equally concerned with the *quality* of judgment and reasoning shown by the subject in the course of the examination Binet was thus using the test situation as an opportunity for a clinical interview—a practice which is becoming increasingly widespread and of increasing importance *in* reports of psychological examinations by present day clinical psychologists

## THE 1908 BINET-SIMON SCALE

Binet and Simon recognized the defects of the first scale They recognized that an improved scale would have to provide more valid norms, based upon a larger and more representative sampling of children at each age, that tests for each age within the limits of the scale would have to be included to achieve finer units of measurement and greater accuracy Their own subsequent investigations and those of other psychologists resulted in a new form of the test, known as the 1908 scale, in which the items are grouped at the appropriate age-levels, from 3 years to 13 years [6]

Age 3

(1) Points to nose, eyes, mouth
(2) Repeats sentences of six syllables
(3) Repeats two digits
(4) Enumerates objects in a picture
(5) Gives family name

Age 4

(1) Knows own sex
(2) Names certain familiar objects shown to him (key, knife, penny)
(3) Repeats three digits
(4) Perceives which is the longer of two lines 5 and 6 cm in length

Age 5

(1) Indicates the heavier of two cubes (3 and 12 grams, 6 and 15 grams)
(2) Copies a square
(3) Constructs a rectangle from two triangular pieces of cardboard, having a model to look at
(4) Counts four coins
(5) Repeats a sentence of ten syllables

Age 6

(1) Knows right and left, indicated by showing right hand and left ear
(2) Repeats sentence of sixteen syllables

---

[6] A Binet and T Simon "Le development de l intelligence chez les enfants," *L Annee Psychologique* Vol 14, 1908, pp 1–94 See also the translation of Binet s publications by E. S Kite *The Development of Intelligence in Children*, Baltimore Warwick and York, 1916

(3) Chooses the prettier in each of three pairs of faces (estbetic comparison)

(4) Defines familiar objects in terms of use

(5) Executes three commissions

(6) Knows own age

(7) Knows morning and afternoon

### Age 7

(1) Perceives what is missing in unfinisbed pictures

(2) Knows number of fingers on each hand and on both hands without counting

(3) Copies a written model ("The little Paul")

(4) Copies a diamond

(5) Describes presented pictures

(6) Repeats five digits

(7) Counts thirteen coins

(8) Identifies by name four common coins

### Age 8

(1) Reads a passage and remembers two items

(2) Adds up the value of five coins

(3) Names four colors red yellow, blue green

(4) Counts backwards from 20 to 0

(5) Writes short sentence from dictation

(6) Gives differences between two objects

### Age 9

(1) Knows the date day of week, day of month, month of year

(2) Recites days of week

(3) Makes change four cents out of twenty in play store transaction

(4) Gives definitions which are superior to use, familiar objects are employed

(5) Reads a passage and remembers six items

(6) Arranges five equal appearing cubes in order of weight

### Age 10

(1) Names the months of the year in correct order

(2) Recognizes and names nine coins

(3) Constructs a sentence in which three given words are used (Paris, fortune, gutter)

(4) Comprehends and answers easy questions

(5) Comprehends and answers difficult questions
(Binet considered item 5 to be a transitional question between ages 10 and 11 Only about one half of the ten year olds got the majority of these correct.)

Age 11

(1) Points out absurdities in statements
(2) Constructs a sentence, including three given words (same as number 3 in age 10)
(3) Gives any sixty words in three minutes
(4) Defines abstract words (charity, justice, kindness)
(5) Arranges scrambled words into a meaningful sentence

Age 12

(1) Repeats seven digits
(2) Gives three rhymes to a word (in one minute)
(3) Repeats a sentence of twenty six syllables
(4) Answers problem questions
(5) Interprets pictures (as contrasted with simple description)

Age 13

(1) Draws the design made by cutting a triangular piece from the once folded edge of a quarto folded piece of paper
(2) Rearranges in imagination the relationship of two reversed triangles and draws results
(3) Gives differences between pair of abstract terms pride and pretension

There are several obvious differences between the 1905 scale and that of 1908 In the former, there are thirty test items, in the latter, fifty-nine The latter does not include the first six items of the 1905 scale which are at the infant level, some other items of the 1905 scale have been eliminated, and many new ones have been added As compared with the 1905 scale, the age-range extends higher in the 1908 scale There are specific groups of items for each age (thus permitting a more accurate rating of individuals), and a greater variety of mental processes is tested

In the 1908 scale, there are also two new and very significant contributions to the theory and practice of mental testing and test construction (1) the tests, after experimentation, were standardized by being grouped into appropriate age levels (Binet's method is explained below), (2) the concept of mental age is employed for the first time [7]

The principal criterion employed by Binet and Simon in the standardization and age placement of tests was this in general, a test was

[7] Although mental age is employed here for the first time, the concept itself had been arrived at by Binet in 1905

placed at the year-level where it was passed satisfactorily by from two thirds to three fourths of a representative group of children of that age. The *ideal* standard was to place a test at a year-level where it was passed by *seventy-five* percent of that age-group. Binet's reason for setting this ideal criterion is a sound one and is made clear by reference to a symmetrical bell-shaped curve, which is approximated by most distributions of intelligence-test scores. The middle 50 percent of the group are most nearly alike, most nearly homogeneous in respect to the abilities being measured, as is obvious from the concentration of these 50 percent within a relatively narrow range, or variation, of scores. Otherwise stated, those individuals constituting the middle 50 percent of the distribution are the *typical* persons of the age group, hence, their test performance should be regarded as typical or normal for their age. If the middle 50 percent of a given age are able to pass a test, then that same test can be passed by the 25 percent who are above the middle group in ability, making a total of 75 who are able to pass the test.

In actual experience, however, it has been practically impossible to devise tests that will exactly satisfy this criterion of 75 percent passing. Fortunately, there are other criteria of validity which are also of primary significance, so that tests are retained if they approximate the 75-percent criterion, and if they demonstrate their value by also satisfying other demands, such as distinguishing between groups of individuals of known ability (mentally deficient, average, and superior), showing appreciable or significant differences between percentages passing at successive age levels, and correlating fairly well with scholastic achievement. These aspects of validation will be more fully discussed in the following chapter, in connection with revisions of the Binet scale.

Binet and Simon standardized their 1908 scale after individual examinations of 203 Paris school children between the ages of 3 and 13 years. While this number is small and would be regarded as inadequate in present-day test-standardization procedures, the fact is that these French pioneers did set a pattern of standardization which is being followed today, with the use of considerable statistical refinement. For in addition to having suggested the criteria already mentioned, they also, in effect, used the symmetrical bell shaped curve as a criterion, though without offering precise numerical values. They stated, simply, that the number of children testing above

age (superior) should equal the number testing below age (inferior), and the number testing at age, or normal, should be greater than the number who rank as either superior or inferior

The mental age, with the 1908 scale, was found this way the subject was credited with the age level at which he passed all the tests To this basic level (now called the basal year') an additional year's credit was added for every five tests passed at higher levels The total was the subject's mental age No credits were given for a fraction of a year, but in the 1911 scale (see below) the calculation of mental age was modified so as to include fractional parts The reader will note that this method of deriving mental age is essentially the same as that used with the American age-scale revisions

In spite of its imperfect standardization, in the 1908 Binet Simon Scale and in the publications concerned with it will be found many of the important concepts and practices which have been employed for about forty years *in the construction* and use of psychological tests

## THE 1911 REVISION OF THE BINET SCALE

The 1908 scale created considerable interest among psychologists in Belgium, Germany, England Italy, Switzerland, and the United States Their interest resulted in a number of valuable applications and evaluations of the Binet scale, accompanied by suggestions for revisions

For the most part, criticisms and suggestions dealt with the age levels at which various items had been placed It is not surprising that, in the first age scale devised to measure intelligence, further and extensive applications and analysis of results should have revealed that a number of the items were misplaced The principal criticism was that the tests at the lower age levels were too easy, whereas those at the higher levels were too difficult, with the result that the former group were rated too high, while the latter were rated too low In other words, standardization of the test had to be improved Binet utilized the suggestions and criticisms of other psychologists, as well as the results of his own continued researches on the 1908 scale, the result being the 1911 revision

Specifically, the major changes incorporated in the 1911 scale were the following four of the tests at the 11 year level were raised to the 12 year level, all 12 year tests were raised to the 15-year level, the

three tests of year 13, plus two new ones, constituted the new adult level Here and there, also, a few tests were placed in either a higher or a lower age level No tests were provided for the 11-, 13-, and 14-year levels [8] In addition to these changes, a fair number of tests found in the 1908 scale were omitted from the 1911 scale because they seemed to depend too much on school learning or on very incidental information

At age levels 3, 4, and 5 the tests are the same as in the 1908 version

Age 6

(1) Distinguishes between morning and afternoon
(2) Defines names of familiar objects in terms of use
(3) Copies a diamond
(4) Counts thirteen *sous*
(5) Distinguishes between pictures of ugly and pretty faces

Age 7

(1) Shows right hand and left ear
(2) Gives description of pictures
(3) Executes three commissions given simultaneously
(4) Gives value of 3 single- and 3 double *sous*
(5) Names four colors red green yellow, blue

Age 8

(1) Gives differences between two objects (from memory)
(2) Counts backwards from 20 to 0
(3) States omissions from unfinished pictures
(4) Knows the date
(5) Repeats five digits

Age 9

(1) Makes change from 20 *sous*
(2) Defines names of familiar objects in terms superior to use
(3) Recognizes all nine [French] coins
(4) Gives months of the year in correct order
(5) Comprehends and answers easy problem-questions

[8] Inasmuch as the rate of mental development appears to decrease appreciably after age 10 it becomes difficult to devise tests which will adequately distinguish between yearly levels This difficulty was encountered also by the authors of the first Stanford Revision of the Binet Simon Scale (1916) but in the second Stanford revision (1937) the authors were able to provide tests at yearly levels between ages 10 and 14 See the next chapter

(2) Intelligence must be measured by testing the higher, complex mental processes rather than relatively simple sensory and motor activities

(3) Intelligence, being a complex, can be tested only by the use of a diversity of materials devised to evaluate the operations of mental processes as an integrated unit, rather than to measure the separate elements that might contribute to the complex functioning of intelligence Though the Binet tests seem to be simple in conception and construction, they actually involve many complex mental activities, memory of several kinds, apperception, free association, orientation in time, language comprehension, ability with numbers, knowledge about common objects, constructive imagination, comparison of concepts, perception of contradictions, understanding of abtract terms, ability to meet novel situations, combining fragments into a meaningful whole

(4) The tests included must be appropriate to the environment of those for whom they are intended

(5) The tests were arranged in the form of a scale, from easiest to most difficult, and groups of tests were placed at appropriate age levels The criterion was, ideally, that a test should be placed at a level where it was passed by three fourths of that age group

(6) The concept of mental age was introduced

(7) The tests must be so standardized that the large middle group of average children (in the curve of distribution) will test "at age"

(8) Other criteria of validity were introduced known groups, scholastic ratings, increase in percentage passing a test at successive age levels

(9) The need of establishing the reliability of a test was recognized, Binet, therefore, made a few reliability studies with his 1911 scale

Binet made not only these contributions he indicated the very extensive uses to which psychological tests could be put in educational, social, vocational, and theoretical problems for he regarded tests as tools for research and for scientific solution of important practical problems Indeed, many of the researches and uses to which tests have since been applied are definitely along lines indicated by Binet, including, among others, the testing of prospective soldiers in order to eliminate the mentally unfit

Binet did not regard his tests as final or as quite satisfactory, he did

not claim that they measured all aspects of personality, he emphasized that they must be supplemented by psychological and educational information derived by other means and from other sources He did claim—and in this he has been supported by extensive subsequent use—that his test, and improved versions that should follow, can provide a very useful and reasonably reliable index of an individual s general intelligence when the tests are administered and interpreted by qualified examiners [9]

---

[9] Alfred Binet died in 1911 His premature death deprived psychology of one of its great pioneers

For a comprehensive study of Binet s psychology see E J Varon *Development of Alfred Binet's Psychology* Psychological Monographs Vol 46 No 3, 1935

# 5.

## EARLY REVISIONS OF THE BINET-SIMON SCALE

### FOUR EARLY REVISIONS

The two most widely known and used adaptations of the Binet scale in the United States are the Stanford revisions of 1916 and 1937 There were, however, four other revisions which, at one time or another, had some currency among psychologists but which today are infrequently used or are chiefly of historical interest In 1908, H H Goddard published a translation of Binet's 1905 scale, and in 1911 he produced, for use in the United States, a revision of Binet's 1908 version Yerkes published revisions in 1915 and 1923, in which the several types of items were grouped as subtests in a point scale (e g , memory span for digits, analogies) instead of being placed at age levels Herring's revision appeared in 1922 and for some years was used as a valuable alternate in place of the 1916 Stanford scale Kuhlmann's three revisions, 1912, 1922, and 1939, were extensive and elaborate in respect to standardization, scoring, and age-range covered Thus it is clear that a significant amount of psychological work had been done prior to the publication of the 1916 Stanford scale and that several investigators continued their research and improvements on Binet's instrument for some years afterwards [1]

[1] Students interested in these historical aspects should consult the following H H. Goddard "A Revision of the Binet Scale," *The Training School Bulletin* (Vineland, N J ), Vol 8, 1911, pp 56–62, F Kuhlmann, *A Handbook of Mental Tests*, Baltimore Warwick and York, 1922, *Tests of Mental Develop-*

**THE STANFORD REVISION OF 1916** [2]

The full name of this test, The Stanford Revision of the Binet-Simon Intelligence Scale is derived from the fact that the revision was made at Stanford University, under the direction of L M Terman The construction of this scale was undertaken for the purpose of providing an instrument that would be adequately standardized and adapted for use in the United States Its acceptance by psychologists and educators is attested by the fact that it was the most widely used individual scale until the revised Stanford Binet appeared in 1937

Although Terman and his collaborators examined approximately 2300 subjects—1700 normal children 200 defective and superior children, and 400 adults—over a period of several years, the revision of the scale below the 14 year level was actually based upon the results obtained with about 1000 native born children in California Each one of these children representing an unselected group of average social status was within two months of his birthday

The 1916 scale includes 90 test items covering an age range from 3 years to 14 years with a group of test items added at the average adult level and another at the superior adult" level Of these 90 test items, 54 were adapted from the 1911 Binet scale 5 from earlier Binet scales 4 from other American tests, and 27 were new additions

Validation The process of selecting the items involved (1) the comments and notes of the examiners including the verbatim responses of the subjects to each test item, and (2) the percentage of subjects passing each test at each age level (as an example, see Table 14) The guiding principle was to secure an arrangement of the tests and a standard of scoring which would make the median mental age of the unselected children of each age group coincide with the median chronological age That is a correct scale must cause the *average* child of 5 years (CA) to test exactly at 5 (MA), the *average* child at 6 to test exactly at 6 etc [3] Or, in terms of the intelligence quo-

---

*ment A Complete Scale for Individual Examination* Minneapolis Educational Test Bureau 1939 R M Yerkes et al *A Point Scale for Measuring Mental Ability* Baltimore Warwick and York 1915 and 1923 J P Herring *Herring Revision of the Binet-Simon Tests* Yonkers N Y World Book 1923 For brief descriptions of these revisions see the 1950 edition of this textbook

[2] L M Terman *The Measurement of Intelligence* Boston Houghton Mifflin 1916

[3] Terman *op cit* p 53

tient employed with this scale, an unselected group of children at each age should yield a median of 100

Before the desired results were secured and this criterion satisfied, it was necessary to prepare three revisions of the scale This involved the elimination of some test items, the shifting of others up or down in age level, and changes in scoring standards [4] "As finally revised," Terman states, 'the scale gives a median intelligence quotient closely approximating 100 for the unselected children of each age from 4 to 14"

The test items above the age of 14 were based on examinations of 30 businessmen, 150 migrating unemployed men, 150 adolescent delinquents and 50 high school students These groups are not a representative cross-section of persons above fourteen years of age in the general population This fact will help make clear why the 1916 scale was found to be unsatisfactory for use with older adolescents and with adults [5] The unsatisfactory quality of the scale at the upper ages was due also to inadequate sampling of abilities

TABLE 14

*Percents Passing Tests Located in Year VI*

1916 Revision [6]

| *Test* | *Ages* 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|
| Right and left | 40 | 50 | 71 | 86 | 95 |
| Mutilated pictures | 27 | 50 | 65 | 87 | 96 |
| Counting 13 pennies | 30 | 46 | 76 | 93 | 86 |
| Comprehension | 25 | 55 | 70 | 86 | 93 |
| Naming four coins | 25 | 47 | 74 | 91 | 95 |
| Repeating 16–18 syllables | 34 | 56 | 69 | 90 | 95 |

[4] A theoretical or ideal percentage of passes for the placement of a test of a given age level was not used Terman states "We had already become convinced that no satisfactory revision of the Binet scale was possible on any theoretical considerations as to the percentage of passes which an individual test ought to show in a given year to be considered standard for that year" (*Op cit.*, p 54 ) Accordingly a "trial and success" method was used in order to get the desired median mental age and IQ at each chronological age level The same practice was followed in standardizing the 1937 revision

[5] In a personal communication Dr Terman states that there were three tentative versions of the scale before the final one was published The businessmen and high-school students were used in making the first tentative placement of tests at average and superior adult levels The other adult groups were then used in subsequent rearrangement of test items

[6] From L M Terman and others *The Stanford Revision and Extension of the Binet Simon Scale for Measuring Intelligence* Educational Psychology Monographs No 18 1917, pp 167–168 (By permission.)

In addition to the criterion of a significant increase in the percent passing a test item at successive ages, the following criteria were used in establishing validity of the 1916 scale

First, in each age group, all the subjects tested were divided into the following three classes those testing below 90 IQ, those testing between 90 and 109, and those testing 110 or above Each test item was then examined to determine whether it was passed by a "decidedly higher" percentage of individuals in the superior IQ group than in the inferior (The term "decidedly higher" was not defined by Terman ) Only those test-items which satisfied this criterion were retained The following data are illustrative

TABLE 15

*Percents Passing Certain Tests Chronological Age Constant*[7]

| | | IQ's | |
|---|---|---|---|
| | *Below 96* | *96–105* | *Above 105* |
| Age 6 Counting thirteen pennies | 40 | 77 | 96 |
| Age 7 Describing pictures | 48 | 52 | 80 |
| Age 8 Giving similarities | 44 | 57 | 83 |
| Age 9 Making change | 39 | 60 | 73 |
| Age 10 Comprehension of problem situations | 25 | 64 | 76 |

Second, after the scale had been developed, the IQ s obtained with 504 school children were compared with their scholastic ratings, as graded by their teachers, on a five point scale, namely, very inferior, inferior, average, superior, very superior Moderate agreement was found between intelligence quotients and school ratings, the coefficient of correlation being 48—close enough so that Terman and his colleagues concluded there was no justifiable 'serious suspicion as to the accuracy of the intelligence scale "

Third the relation between IQ and grade progress was studied for the children on whom the scale was standardized A 'fairly high" correlation was found, but there were also some "astounding disagreements' inasmuch as a given mental age level was found in a wide range of grades For example a mental age of nine was found in all grades from 1 to 7 Terman states however, 'When the data were examined it was found that practically every child whose grade failed

[7] From Terman and others *ibid* p 133 (By permission.)

to correspond fairly closely with his mental age was either exceptionally bright or exceptionally dull Those who tested between 96 and 105 IQ [the average children] were never seriously misplaced in schools "[8]

**Reliability.** Following its publication, the Stanford-Binet was subjected to numerous studies in order to determine its reliability by the method of self-correlation The correlation coefficients, which in such studies will vary with the size and constitution of the experimental group, ranged from about 80 to 95 Such coefficients are regarded as highly satisfactory indexes of reliability

**The 1916 Scale.** The reader will have noted that no essentially new concepts or principles have been added in the 1916 Stanford-Binet scale, as compared with Binet's own Terman and his colleagues did, however, extend, refine, and adapt the Binet scales, so that the 1916 revision was a better standardized, hence more valid and reliable, instrument

The complete list of tests of the 1916 Stanford Binet follows[9] Throughout the scale, those items designated "Al," instead of being numbered, are alternates, to be used in place of one of the numbered items, where the examiner, for any reason, believes a numbered item to be inappropriate

Age 3

(1) Points to parts of body
(2) Names familiar objects
(3) Enumerates objects in pictures
(4) Gives sex
(5) Gives last name
(6) Repeats six to seven syllables
(Al) Repeats three digits

Age 4

(1) Compares lengths of lines
(2) Discriminates between geometric forms
(3) Counts four pennies
(4) Copies a square
(5) Comprehends and solves problem situations
(6) Repeats four digits
(Al) Repeats twelve to thirteen syllables

[8] Terman *The Measurement of Intelligence* p 74
[9] Reproduced by permission of Houghton Mifflin.

### Age 5

(1) Compares weights
(2) Names familiar colors
(3) Makes esthetic comparisons of paired drawings of faces
(4) Defines common words use or better
(5) Puts together a divided triangle
(6) Carries out three commissions
(Al ) Gives own age

### Age 6

(1) Knows right from left
(2) Perceives missing parts in pictures
(3) Counts thirteen pennies
(4) Comprehends and solves problem situations
(5) Identifies coins
(6) Repeats sixteen to eighteen syllables
(Al ) Knows morning from afternoon

### Age 7

(1) Knows number of fingers on each and both hands
(2) Describes pictures
(3) Repeats five digits
(4) Ties a bow knot
(5) Gives differences between paired objects
(6) Copies a diamond
(Al 1) Names days of week in correct order
(Al 2) Repeats three digits backwards

### Age 8

(1) Traces path to be followed in a systematic search for a lost object in a field
(2) Counts backwards from 20 to 1
(3) Comprehends and solves problem situations
(4) Gives similarities between two things
(5) Defines names of objects in terms superior to use
(6) Defines twenty words from a vocabulary list
(Al 1) Identifies six coins
(Al 2) Writes short sentence from dictation

### Age 9

(1) Gives date day of week month day of month year
(2) Discriminates between weights 3 6 9 12 15 grams
(3) Makes change in small amounts
(4) Repeats four digits backwards
(5) Makes up a sentence including three given words
(6) Gives rhymes to three words

(Al 1) Names the months of the year
(Al 2) Gives total value of a group of one cent and two cent post age stamps

### Age 10

(1) Defines thirty words from vocabulary list
(2) Detects absurdities in statements
(3) Reproduces two designs from memory
(4) Reads a short passage and reproduces content
(5) Comprehends and solves problem situations
(6) Names any sixty words by free association
(Al 1) Repeats six digits
(Al 2) Repeats twenty to twenty two syllables
(Al 3) Fits rectangular blocks into form board

### Age 12

(1) Defines forty words from vocabulary list
(2) Defines abstract words
(3) Traces a path in systematic search (same problem as in year 8 but a superior plan is required here)
(4) Rearranges dissected sentences into meaningful sentences
(5) Interprets fables
(6) Repeats five digits backwards
(7) Interprets pictures
(8) Gives similarities between three things

### Age 14

(1) Defines fifty words from vocabulary list
(2) Discovers a rule in a paper folding test (induction test)
(3) Gives differences between a president and a king
(4) Integrates given facts and arrives at a conclusion concerning them
(5) Solves arithmetical reasoning problems
(6) Reverses hands of clock in imagination and gives the hour
(Al ) Repeats seven digits

### Average Adult

(1) Defines sixty five words from vocabulary list
(2) Interprets fables
*(3) Gives differences between abstract words*
(4) Solves problem of number of enclosed boxes (boxes within boxes) when shown only the large outside box
(5) Repeats six digits backwards
(6) Perceives the pattern of a code and uses it
(Al 1) Repeats 28 syllables
(Al 2) Comprehends problems involving physical relations

Superior Adult

(1) Defines seventy five words from vocabulary list
(2) Visualizes, imaginally, and draws appearance of a folded and cut piece of paper
(3) Repeats eight digits
(4) Repeats thought of a passage heard
(5) Repeats seven digits backwards
(6) Solves problems involving 'ingenuity'

The Scoring Method. Each age level from 3 years through 10, it will be noted, has six test items (plus the alternate which may replace one of the six) Each of these carries credit of two "months," so that the tests in each of the age levels provide a year's increment in mental age

There are no tests at the eleven-year level, the reason being that the authors of the scale were, apparently, unable to devise tests that would indicate a one-year difference at this stage of mental development This gap in the scale, it is believed, is due to the slowing down of mental development, thus decreasing the annual increments and making it more difficult to measure those increments by means of the then-available test items Since the eight test items at age twelve cover a span of two years, each one carries a credit of three months in order to yield an average mental age of twelve, when added to the ten year level

The same explanation applies to the six test items at age 14, each of which gives a credit of four months toward mental age score [10]

The tests and credits at the "average adult" level were devised so as to provide a median value mental age of 16 Yet each of the six tests at average adult level carries a credit of five months, so that a person who passed all of them would get a mental age of 16 5 years In his volume describing the standardization of the 1916 revision, in explaining the limit for "average adult," Terman states that his data on mental ages of 62 adults, including 30 businessmen and 32 high school pupils, who were over 16 years of age, show "   that the middle section of the graph [of the distribution] represents the 'mental ages' falling between 15 and 17 This is the range we have designated as the 'average adult' level " [11]

Those persons having mental ages above seventeen were designated

[10] The 1937 revision provides a group of tests at age 11 and another at 13
[11] Terman, *The Measurement of Intelligence* p 55

as "superior adults," the possible maximum mental age on this scale being 19 5 (Six tests, six months' credit each, added to the maximum of 16 5 attainable at the "average adult" level )

The method of scoring the 1916 scale is this The examiner goes *down* in the test until the level is reached where the subject passes all items This is called the "basal year " The examiner then proceeds *upward* in the scale until the level is reached where the subject fails all items This level is called the "terminal year " As already stated, each test item carries specified credit, in terms of months, contributing to the mental age score These credits are added to the age value of the basal year, the total is the mental age For example assume that in a given instance the basal year is 6, three test items are then passed at the 7-year level, giving additional credit of 6 months, two are passed at the 8-year level, giving further credit of 4 months, all are failed at the 9 year level Thus, this subject's mental age is 6 years, 10 months

**Distribution of IQ's.** A second score obtained with the 1916 Stanford-Binet is the intelligence quotient, the calculation and general nature of which have already been explained Terman not only found the IQ for each individual examined, but he analyzed the distribution of intelligence quotients obtained by the persons on whom the scale was standardized

Taking those subjects between the ages of 5 and 14 years, the distribution was found to be the following

TABLE 16

*Distribution of IQ s of 905 Unselected Children, Ages 5–14 Years*[12]

| IQ | Percent of total |
|---|---|
| 56– 65 | 0 33 |
| 66– 75 | 2 3 |
| 76– 85 | 8 6 |
| 86– 95 | 20 1 |
| 96–105 | 33 9 |
| 106–115 | 23 1 |
| 116–125 | 9 0 |
| 126–135 | 2 3 |
| 136–145 | 0 55 |

[12] From Terman and others, *op cit* p 40 (By permission.)

The standard deviation (S D ) of this distribution is about 12 points (Compare this with the S D of 16 points of the 1937 revision )

This distribution, being a fairly symmetrical one, showed that the scale did differentiate between the several levels of mental capacity of the persons examined, at least, so far as concerns the mental processes being tested It therefore strengthened the belief of Terman and many others that the 1916 Stanford-Binet had considerable validity

Another method used to represent the frequency with which different degrees of intelligence occur was to indicate the percentage of subjects at and above or at and below a given IQ, thus.

TABLE 17

*Percentage Distribution of IQ's Stanford Binet Scale, 1916*[13]

| | | | | | |
|---|---|---|---|---|---|
| The lowest | 1% | go to | 70 | or | below |
| " " | 2% | " " | 73 | " | " |
| " " | 3% | " " | 76 | " | " |
| " " | 5% | " " | 78 | " | " |
| " " | 10% | " " | 85 | " | " |
| " " | 15% | " " | 88 | " | " |
| " " | 20% | " " | 91 | " | " |
| " " | 25% | " " | 92 | " | " |
| " " | 33⅓% | " " | 95 | " | " |
| The highest | 1% | reach | 130 | or | above |
| " " | 2% | " | 128 | " | " |
| " " | 3% | " | 125 | " | " |
| " " | 5% | " | 122 | " | " |
| " " | 10% | " | 116 | " | " |
| " " | 15% | " | 113 | " | " |
| " " | 20% | " | 110 | " | " |
| " " | 25% | " | 108 | " | " |
| " " | 33⅓% | " | 106 | " | " |

Although the foregoing percentages above or below certain IQ levels are not fixed or identical for all tests (e g , the distribution for the 1937 Stanford-Binet is not identical with this), a table such as this is significant in that it provides one means of determining an individual's relative status in respect to the psychological processes being

[13] From Terman, *The Measurement of Intelligence*, Boston Houghton Mifflin, 1916, p 78 (By permission.)

measured, for, as already explained, the IQ is an index having educational and clinical connotations

On the basis of the distribution of intelligence quotients obtained with the 1916 revision, Terman also suggested the following classification

TABLE 18

*Suggested Classification of IQ s Stanford Binet Scale, 1916* [14]

| *IQ* | *Classification* |
|---|---|
| Above 140 | Near' genius or genius |
| 120–140 | Very superior intelligence |
| 110–119 | Superior intelligence |
| 90–109 | Normal or average, intelligence |
| 80–89 | Dullness |
| 70–79 | Borderline deficiency |
| Below 70 | Definite feeblemindedness |

This classification is reproduced because it has been widely used and because the reader should be familiar with its source Unfortunately, however, such classifications or labels have frequently been used uncritically, the erroneous assumption having been that there was some quality inherent in the classification itself that warranted the designation of "genius," or "superiority," or "dullness," etc It has already been pointed out, for example, that not all persons of very high IQ s have original, creative mentalities, yet these are among the traits of the "genius ' The particular IQ intervals and the names attached to them result from the judgment of the specialist making the classification Some classifiers might choose to place the lower limit of near genius or genius" at 150 IQ, or the upper limit of ' feeblemindedness ' at 50 or 60 In short, tables of IQ classifications are essentially statistical

Regardless of size of intervals or of their names, a classification is useful chiefly as a convenient device for purposes of research and analysis of data It should not be used merely to label and pigeonhole an individual who has been examined and for whom an intelligence quotient has been obtained The trend is away from stating test results in terms of MA and IQ alone, the trend is in the direction of evaluation of individual performances

[14] From Terman, *op cit*, p 79 (By permission )

**Adult Mental Age and Adult Intelligence Quotient.** Since the 1916 Stanford Binet includes tests at the levels of average adult and superior adult it was necessary to make provisions for the calculation of adult mental ages and intelligence quotients These, however, present special problems

We have already quoted Terman's reason for locating average adult performance in the mental age range of 15 to 17, with the assumed midpoint at 16 years If this is correct, it means that the test performance of the average adult is equal to that of the average 16-year-old individual Otherwise stated, it means that in the case of an *average* adult his maximum level of measured intelligence is reached at the age of 16 and that there are no increments thereafter Terman states that "  in so far as it can be measured by tests now available, [intelligence] appears to improve but little after the age of 15 or 16 Although this point [at which intelligence attains its final development] is not exactly known it will be sufficiently accurate for our purposes to assume its location at 16 years "[15] Thus until the process of decline sets in *the average adult continues to have a mental age of 16* according to the 1916 Stanford Binet

On the basis of this assumption, then, in the calculation of an IQ for a person who is 16 years of age, or older, the denominator in the formula ($IQ = \frac{MA}{CA}$) is always 16 Otherwise, if his actual CA were used, he would *appear* to be getting rapidly less and less intelligent with the succeeding years For example, an average individual at the age of 16 will have an IQ of 100 (16/16) At the age of 18, he should still have an IQ of 100 even though, according to the tests being used, there has been no further measurable development of mental capacity, for the formula will still be 16/16 Now if in the case of this same individual, we had continued using his actual CA as the denominator, his IQ at age 18 would be shown as about 89 (16/18), at age 20, it would be shown as 80 (16/20), and so on, while as a matter of fact there would ordinarily have been no such decline Thus, by using the denominator of 16 in the IQ formula for all persons above age 16, if a person of 20 years and one of 60 years have the mental age of 16 years each then each will be given an IQ of 100

The reader has also noted that it is possible to attain mental ages

[15] *Op cit* p 140

above 16 on the 1916 revision, the maximum being 19 5 years at the level of superior adult If the definition of mental age is borne in mind, it will be apparent at once that a mental age rating which is higher than that of the *average* adult has a new and specialized meaning It cannot have the same meaning as the term, mental age, does ordinarily A mental age is defined as the level of mental development of the *average* or *typical group* of persons at that same chronological age Thus, an MA of 10 represents the test performance and mental level of a group of average children of chronological age 10 Hence, if we assume that *average* or typical adults reach a mental age of 16, then to speak of a "mental age' above 16 is to introduce a new concept, for these latter "mental ages" are not derived from the performance or norms of average or typical persons They are theoretical and hypothetical indexes devised to enable us to indicate higher than average mental levels and higher than average intelligence quotients [16] Thus, when a higher-than average adult "mental age" is used, it is essential that the user be aware of the fact that a new and different concept is being employed

The fact that the highest possible "mental age" that can be attained on this test is 19 5 years means that the highest IQ an adult can get is about 122 (19 5/16) This maximum reveals a serious inadequacy of the 1916 revision at the higher levels What, for example, happens to the IQ of the 10-year-old child who has a mental age of 15, and an IQ of 150 (15/10)? Obviously, to maintain that IQ of 150 at age 16 or older, he must be able to attain a mental age of 24 (24/16), yet the scale permits a maximum MA of only 19 5, with an IQ of 122 The same would be true for this subject after age 16

**Criticisms of the 1916 Stanford-Binet.** Experience with the 1916 Stanford Binet demonstrated that it was inadequate as a measure of adult mental capacity In fact, experience showed that its usefulness was restricted to ages between five and fourteen years, the range between five and ten years being the most satisfactory

[16] In their volume describing the 1916 revision (*op cit*) Terman and his collaborators do not report how they arrived at their mental age levels above those of average adult. In their 1937 revision this has been done by extrapolation and by providing for a distribution of adult IQ s which should correspond to distributions for pre adult levels

This revision was also criticized on several other grounds First, since the scale was finally standardized on the basis of results obtained with approximately 1000 native born children in California, its use with *all* groups of children in *all* parts of the United States seemed to many educators and psychologists to be a practice of doubtful validity, for it was held that the 1000 California subjects were not necessarily representative of the child population of this country There is merit in the criticism, yet it must be recognized that this scale proved to be very useful in many parts of the United States, when employed and interpreted by examiners who were familiar with its assumptions and construction, and who, at the same time, were familiar with the backgrounds of the subjects they were examining

Second, the scale was criticized as being much too heavily weighted with verbal and abstract materials, thus penalizing the individual who, for whatever reason, had been handicapped in developing his "verbal intelligence" through the medium of the English language Terman's reply to this criticism was that intelligence at the verbal and abstract levels is the highest form, the *sine qua non,* of mental ability Indeed, he defined intelligence as the ability to deal with abstract terms, and to do conceptual thinking

This criticism of the scale was warranted, nevertheless, for children who are handicapped by lack of opportunity to acquire and develop the use of the English language are at a serious disadvantage and get spuriously low ratings on psychological tests which emphasize "verbal intelligence " Such children would include (1) those who have developed in homes where only a foreign language is spoken, (2) those who are handicapped by serious visual or auditory defects, (3) those handicapped by sensory anomalies (reversals, inversions, mirror-writing, poor sound discrimination) which seriously interfere with their learning to read, (4) those who are too young—say, below age four or five—to be tested adequately by means of verbal materials almost exclusively

Third, the 1916 scale was found to be defective at some points in respect to procedures in administering and scoring, thus detracting from its objectivity and from the comparability of results obtained by different examiners

In view of these criticisms, it was to be expected that other scales should be developed, especially those of the "performance" and "non-

verbal' type, which would obviate or minimize the second criticism These will be presented and discussed in a later chapter

It was to be expected, also, that the 1916 Stanford-Binet itself should undergo revision in the light of experience, criticism, and accumulated data Such a revision, begun about ten years after the original Stanford-Binet appeared, was published in 1937.

# 6.

## THE 1937 REVISION OF THE STANFORD-BINET SCALE

### DESCRIPTION OF THE 1937 SCALE

This scale differs from that of 1916 in many details, but it does not differ in its essential and basic conceptions.[1] As the authors themselves state, "The revision utilizes the assumptions, methods, and principles of the age scale as conceived by Binet." The authors, however, do regard it as a better standardized and more useful scale than its predecessors. The principal differences and modifications follow.

The 1937 scale has two equivalent forms (L and M), each of which contains 129 test items, as compared with the 90 items in the first Stanford Binet. Items that proved unsatisfactory in the original were eliminated, and new ones were added.

The 1937 scale extends downward to the level of age 2, and upward through three levels of 'superior adult' (known as Superior Adult I, II, and III), thus increasing its usefulness.

The levels below age 5 and those above age 14 have been more carefully and validly standardized.

Scoring standards and instructions for administering the tests were improved.

[1] L. M. Terman and M. A. Merrill, *Measuring Intelligence*, Boston, Houghton Mifflin, 1937. See also R. Pintner and others, *Supplementary Guide for the Revised Stanford Binet Scale (Form L)*, Applied Psychology Monographs, No. 3, 1944. This monograph presents, for purposes of comparison, a collection of responses to test items but does not duplicate those provided in the Terman Merrill manual.

From the age of 2 to age 5, this scale provides groups of test items at half-year intervals. Thus more accurate and more highly differentiating test results are obtainable. The half-yearly intervals are possible because the rate of mental growth is most rapid in the earlier years and, therefore, the more rapid periodic increments are susceptible to testing

Groups of tests are provided at ages 11 and 13, whereas there were none at these levels in the 1916 scale for reasons already stated.

Although the 1937 scale, like that of 1916, is predominantly verbal in character, it does provide more performance and other nonverbal materials at the earlier age levels, especially through the age of four years. Performance materials are those with which the subject has to *do* something; for example, building a pattern or making a design with blocks, or filling in a form board with the variously shaped blocks Other nonverbal materials include such activities as copying a geometric figure, completing the picture of a man, discriminating between forms, etc. In all of these, of course, verbal ability is a factor to the extent that verbal directions must be understood In these tests, verbal ability can also operate as a factor if the subject is familiar with the names of the objects or geometric figures and is thus facilitated in his manipulation or classification of them

The 1937 scale has been standardized on a more carefully chosen and much more extensive group of subjects. The base of the standardization population was broadened, and its component members are regarded as more representative of the population [2] *But only American-born white subjects were used in the standardization of this scale*, the total number being approximately 3000. The subjects were chosen from eleven states in several widely separated areas of the country, and an effort was made to have the subjects from homes which, occupationally and socially, would be representative of the population at large in the United States.

## VALIDATION

The test items were chosen on the basis of their validity, ease and objectivity of scoring, economy of time in administering, interest to the subjects, and need for variation in types of materials.

[2] Up to the age of 5, the number of subjects used was 100 to each half year level from 6 to 14 years of age, 200 at each year; from 15 to 18 years, 100 at each year.

Of the foregoing, *validity* is of primary significance In this revision, a criterion of basic importance in judging validity of test items was increase in percentage of successful performance with increasing age This criterion was applied in two ways first, by requiring an appreciable increase in the percent passing a given item in successive ages (as in the 1916 scale), and second, by finding "a weight based on the ratio of the difference to the standard error of the difference between the mean age (or mental age) of subjects passing the test and of subjects failing it "[3] Stripped of its statistical terminology, what this quotation means is that the difference between the average age (chronological or mental) of subjects passing an item, on the one hand, and the average age of subjects failing that item, on the other hand, must be statistically significant This is essentially an "age criterion " In this connection see Table 19

TABLE 19

Percent Passing Test Items Located in Year VI,[4] Form L

| | Ages | | | | | | |
|---|---|---|---|---|---|---|---|
| *Item* | *4* | *4½* | *5* | *5½* | *6* | 7 | 8 |
| 1 | 3 | 15 | 36 | 50 | 67 | 89 | 97 |
| 2 | 11 | 29 | 44 | 55 | 70 | 86 | 95 |
| 3 | 11 | 26 | 46 | 53 | 69 | 86 | 96 |
| 4 | 3 | 11 | 43 | 48 | 71 | 94 | 96 |
| 5 | 16 | 29 | 47 | 51 | 73 | 94 | 95 |
| 6 | 26 | 44 | 52 | 61 | 81 | 91 | 93 |

A second criterion of major importance in the retention of an item was its correlation with the *total scores* of the individuals of the age level at which the test item is located Table 20 presents the distribution of correlation coefficients (biserial) for both forms L and M

The calculated median of the coefficients for form L is approximately 69, the middle 50 percent of the coefficients fall between approximately 51 and 73 The range for the whole set of coefficients is from 28 (memory for designs, year 11) to 89 (abstract words, year 11, and vocabulary, year 14)

[3] Terman and Merrill, *op cit* p 9 See also V V Fleming A Study of the Subtests in the Revised Stanford Binet Scales Forms L and M," *Journal of Genetic Psychology* Vol 64 1944 pp 3–36

[4] From Q McNemar *The Revision of the Stanford Scale* Boston Houghton Mifflin, 1942, p 92 (By permission )

On from M, the calculated median coefficient is approximately .64; the middle 50 percent of the coefficients fall between approximately .51 and .71. The range for this whole set of coefficients is from .27 (memory for stories, year 13) to .91 (abstract words, year 13)

Of the 258 coefficients, 201, or very nearly 78 percent, are .50 or higher. This fact and the data presented in a later section in this chapter (*Analysis of Functions Tested*) provide strong evidence that the Stanford-Binet Scale measures "general ability" by means of test items that have psychological functions in common to a high degree.

TABLE 20

*Distribution of Correlation Coefficients (Biserial) for Each Item with Total Scores*[3]

| *r* | *Frequency Form L* | *Frequency Form M* |
|---|---|---|
| 20– | 1 | 2 |
| 30– | 5 | 7 |
| 40– | 21 | 21 |
| 50– | 28 | 23 |
| 60– | 32 | 42 |
| 70– | 34 | 23 |
| 80– | 8 | 9 |
| 90– | | 2 |
| N | 129 | 129 |

After selection of the tests that were to be used, one other empirical procedure was employed in locating each test item at an appropriate age level. The items were rearranged until it was found that they would yield for each group of subjects a mean mental age that was identical with their mean chronological age, giving a mean IQ as close as possible to 100. Six successive revisions were necessary to achieve this for Form L. Then, ". . . it was possible to achieve at once an equally good result with Form M by arranging its tests so as to match those of Form L at each age level with respect to difficulty, validity, and shape of curves of percents passing by age."[4]

RELIABILITY

Comparing IQ's obtained with Forms L and M, Terman and Merrill report reliability coefficients of correlation from .90 to .98

[3] Based upon data in Q. McNemar, *op. cit.*, Tables 53 and 54.

[4] Terman and Merrill, *op. cit.*, p. 23.

The highest coefficients were found for IQ's below 70 (.98), the lowest for IQ's above 130 (.90), and intermediate coefficients for IQ's near 100 (.92). Age levels above six years showed greater reliability (.93) than those below six (.88), calculated for separate age-groups.

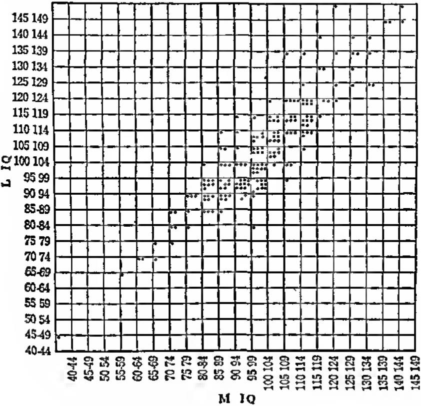



FIG 6.1 *Scatter Chart of Correlation between Form L and Form M IQ's at Chronological Age 7. r = .91. From Terman and Merrill, op. cit., p. 45. (By permission.)*

These coefficients are of the same general order as those found by others in subsequent studies. (See Figure 6.1.)

## DETERMINING MENTAL AGE AND INTELLIGENCE QUOTIENT

The scoring method is exactly the same in principle as that used with the 1916 scale for the determination of mental age and intelligence quotient. There are, however, a few differences in the details.

Whereas the maximum mental age attainable on the 1916 Stanford-Binet was 19 years and 6 months, the maximum on the 1937 revision is 22 years and 10 months. It will be recalled that with the first Stanford-Binet scale, a maximum CA of 16 was used in the denominator to determine the IQ of an individual of 16 years of age or older In the 1937 scale, the maximum CA in the denominator is 15. Thus, the highest possible IQ attainable by a subject who is 15 years of age or older is 152 (22 ⁴/15).

In the superior-adult levels, the test items were selected and their credits allotted (in terms of months) in such a manner as to make the IQ distribution of ". . . the older subjects resemble closely those of the younger, as presumably should be the case on an ideal scale "[7]

TABLE 21

*Correction of CA Divisor, 1937 Stanford Binet Scale*[8]

| Actual CA | Corrected CA Divisor |
|---|---|
| 13–0 | 13–0 |
| 13–3 | 13–2 |
| 14–0 | 13–8 |
| 14–6 | 14–0 |
| 15–0 | 14–4 |
| 15–6 | 14–8 |
| 16–0 | 15–0 |

In order to achieve this desired goal, it was necessary for the authors to make adjustments in the denominator of the IQ formula, beginning at the age of 13 years and 2 months. The reason given for this adjustment is that it is extremely difficult, perhaps impossible, to escape the effects of selection of subjects at the upper ages in standardizing a scale. The selection generally is such as to include the average range and the higher mental levels but not the lowest, since the less intelligent individuals tend to leave school earlier than the others. Hence, the *norms* of test-performance of the older groups, it is argued, tend to be higher than they should be for an unselected sampling. These higher norms, in turn, tend to reduce the intelligence quotients of subjects in the older groups. It is this effect that the authors of the scale sought to correct by means of adjusting the denominator.

[7] *Ibid*, p 30

[8] From Terman and Merrill, *op cit*, p 31 (By permission.)

While Terman and Merrill believe they minimized the effects of selection they were not wholly eliminated Therefore after a 'trial-and success process directed toward making IQ distributions of older age groups resemble closely those of younger groups the procedure adopted was this from age 13 to age 16, the cumulative dropping of one out of every three additional months of chronological age and all of it after 16 In substance this practice is tantamount to saying that

TABLE 22

*IQ Means Adjusted for 1930 Census Frequencies of Occupational Groupings* [a]

*Composite of Forms L and M Stanford Binet*

| *Age* | *N* | *Raw* | *Smoothed* |
|---|---|---|---|
| 2 | 76 | 102 1 | |
| 2½ | 74 | 104 7 | 103 3 |
| 3 | 81 | 103 2 | 104 1 |
| 3½ | 77 | 104 3 | 102 2 |
| 4 | 83 | 99 2 | 101 6 |
| 4½ | 79 | 101 2 | 100 8 |
| 5 | 90 | 101 9 | 100 4 |
| 5½ | 110 | 98 2 | 100 0 |
| 6 | 203 | 100 0 | 99 8 |
| 7 | 202 | 101 2 | 100 8 |
| 8 | 203 | 101 1 | 102 0 |
| 9 | 204 | 103 6 | 102 7 |
| 10 | 201 | 103 5 | 103 0 |
| 11 | 204 | 101 9 | 102 2 |
| 12 | 202 | 101 2 | 101 6 |
| 13 | 204 | 101 8 | 101 0 |
| 14 | 202 | 100 0 | 101 3 |
| 15 | 107 | 102 0 | 101 3 |
| 16 | 102 | 101 8 | 103 3 |
| 17 | 109 | 103 2 | 103 8 |
| 18 | 101 | 106 3 | |

average adult mental age, on the 1937 scale, is 15 Table 21 gives a few examples

When therefore, this scale is used with subjects who are older than thirteen years, it is necessary to refer to the full correction table provided in the manual Or the examiner may use the tables of IQ s provided in the manual in which the necessary adjustments have already been made

[a] From Terman and Merrill *op cit* p 36 (By permission )

**DISTRIBUTION OF IQS**

The mean IQ s for the subjects used in the standardization are slightly above 100 But this the authors say, is due to an intentional adjustment to allow for the somewhat inadequate sampling of subjects in the lower occupational classes The adjustment was made by dividing the subjects into seven groups according to the occupation of the fathers at each age level the mean IQ was computed sepa

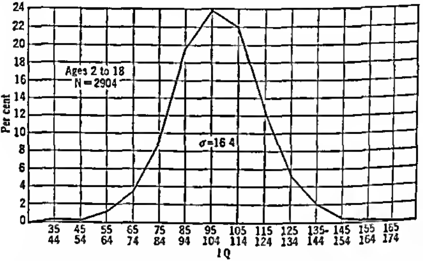



FIG 6 2 Distribution of Composite L-M IQ s of Standardization Group Terman and Merrill *Measuring Intelligence* Boston Houghton Mifflin p 37 (By permission )

rately for each of the seven groups These means at each age level were given a weight according to the occupational frequencies of each group as shown by the 1930 census The weighted means were then combined into a composite mean for each age level from 2 to 18 years as shown in Table 22 The same data are represented graphically in Figure 6 2

In determining the equality and comparability of IQ s from age to age it is necessary not only that the means be very much the same (ideally identical) but that the *variations* be the same at all age levels If the differences between the variations of the age groups are

large, then the same numerical IQ will have different significance at different chronological ages

Consider the following hypothetical instance Suppose a given test of mental ability yields the following results

| *Chronological Age* | *Mean IQ* | *Standard Deviation* |
|---|---|---|
| 10 | 100 | 14 |
| 11 | 100 | 20 |

Accordingly, a 10-year-old child having an IQ of 86 (that is, one standard deviation below the mean) would have a percentile rating of approximately 16—which it will be recalled means that this child surpasses about 16 percent of his age group Now, according to the foregoing data a child of 11 years whose IQ is 80 (likewise one standard deviation below the mean of *his* group) would also have a percentile rating of about 16 in spite of the fact that his intelligence quotient is six points below that of the 10 year old in question While this difference of six points may make little practical difference in the clinical and educational treatment of these children it is necessary to be familiar with the implications of differences in variations

Another aspect of the problem is this Taking our hypothetical case of the 10 year-old boy, if he is to have at age 11 the same *relative rank* he held at age 10 his IQ will have to drop from 86 to 80 and if this happens the change will give the appearance of lack of IQ constancy, but if he maintains his 86 IQ at age 11 then his percentile rank will be approximately 24 Here again the difference in percentile ranks might have little practical significance but insight into the effects of differences in group variations will help explain some instances of lack of close agreement in repeated mental measurements *Furthermore the professional worker who uses mental tests must know the standard deviations of the instruments he employs in order to make as accurate an evaluation of his results as possible*

IQ variability in relation to age, as found in the standardization of the 1937 Stanford Binet is shown in Table 23 Inspection of this table shows significant fluctuation in standard deviations of the several age groups especially between the extremes namely, 12 5 (S D ) at age 6, 20 6 (S D ) at age 2½, and 20 0 (S D ) at age 12 It will be noted, too that the standard deviations fluctuate around 16 and 17 as a median value The standard deviation of the composite IQ s (Forms L and M) is 16 4 for the entire standardization group of subjects

DISTRIBUTION OF IQ S

The mean IQ s, for the subjects used in the standardization, are slightly above 100 But this, the authors say, is due to an "intentional adjustment to allow for the somewhat inadequate sampling of subjects in the lower occupational classes " The adjustment was made by dividing the subjects into seven groups, according to the occupation of the fathers, at each age level the mean IQ was computed sepa

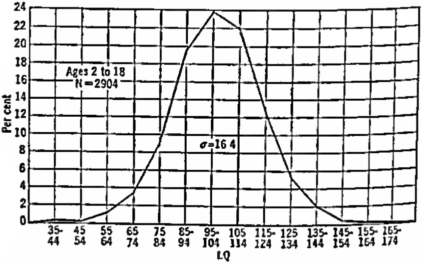



FIG 6 2 *Distribution of Composite L-M IQ s of Standardization Group* Terman and Merrill *Measuring Intelligence* Boston Houghton Mifflin p 37 (By permission )

rately for each of the seven groups These means at each age level were given a weight according to the occupational frequencies of each group as shown by the 1930 census The weighted means were then combined into a composite mean for each age level from 2 to 18 years, as shown in Table 22 The same data are represented graphically in Figure 6 2

In determining the equality and comparability of IQ s from age to age it is necessary not only that the means be very much the same (ideally, identical), but that the *variations* be the same at all age levels If the differences between the variations of the age groups are

large then the same numerical IQ will have different significance at different chronological ages

Consider the following hypothetical instance Suppose a given test of mental ability yields the following results

| *Chronological Age* | *Mean IQ* | *Standard Deviation* |
|---|---|---|
| 10 | 100 | 14 |
| 11 | 100 | 20 |

Accordingly, a 10 year-old child having an IQ of 86 (that is, one standard deviation below the mean) would have a percentile rating of approximately 16—which, it will be recalled, means that this child surpasses about 16 percent of his age group Now, according to the foregoing data, a child of 11 years whose IQ is 80 (likewise one standard deviation below the mean of *his* group) would also have a percentile rating of about 16, in spite of the fact that his intelligence quotient is six points below that of the 10-year-old in question While this difference of six points may make little practical difference in the clinical and educational treatment of these children, it is necessary to be familiar with the implications of differences in variations

Another aspect of the problem is this Taking our hypothetical case of the 10-year-old boy, if he is to have at age 11 the same *relative rank* he held at age 10, his IQ will have to drop from 86 to 80, and if this happens, the change will give the appearance of lack of IQ constancy, but, if he maintains his 86 IQ at age 11, then his percentile rank will be approximately 24 Here, again, the difference in percentile ranks might have little practical significance, but insight into the effects of differences in group variations will help explain some instances of lack of close agreement in repeated mental measurements *Furthermore the professional worker who uses mental tests must know the standard deviations of the instruments he employs in order to make as accurate an evaluation of his results as possible*

IQ variability in relation to age, as found in the standardization of the 1937 Stanford Binet is shown in Table 23 Inspection of this table shows significant fluctuation in standard deviations of the several age groups, especially between the extremes namely, 12 5 (S D ) at age 6, 20 6 (S D ) at age 2½, and 20 0 (S D ) at age 12 It will be noted, too, that the standard deviations fluctuate around 16 and 17 as a median value The standard deviation of the composite IQ's (Forms L and M) is 16 4 for the entire standardization group of subjects

In respect to the fluctuations in IQ variability, the authors state ' Notwithstanding our strenuous efforts to correct for errors of sampling complete success is hardly to be expected, and a considerable degree of irregular fluctuation in the found magnitudes of IQ variability from age to age could reasonably be attributed to these sources of error Since inspection of the values reveals no

TABLE 23

*IQ Variability in Relation to Age* [10]

*Stanford Binet*

| CA | N | S D Form L | S D Form M |
|---|---|---|---|
| 2 | 102 | 16 7 | 15 5 |
| 2½ | 102 | 20 6 | 20 7 |
| 3 | 99 | 19 0 | 18 7 |
| 3½ | 103 | 17 3 | 16 3 |
| 4 | 105 | 16 9 | 15 6 |
| 4½ | 101 | 16 2 | 15 3 |
| 5 | 109 | 14 2 | 14 1 |
| 5½ | 110 | 14 3 | 14 0 |
| 6 | 203 | 12 5 | 13 2 |
| 7 | 202 | 16 2 | 15 6 |
| 8 | 203 | 15 8 | 15 5 |
| 9 | 204 | 16 4 | 16 7 |
| 10 | 201 | 16 5 | 15 9 |
| 11 | 204 | 18 0 | 17 3 |
| 12 | 202 | 20 0 | 19 5 |
| 13 | 204 | 17 9 | 17 8 |
| 14 | 202 | 16 1 | 16 7 |
| 15 | 107 | 19 0 | 19 3 |
| 16 | 102 | 16 5 | 17 4 |
| 17 | 109 | 14 5 | 14 3 |
| 18 | 101 | 17 2 | 16 6 |

marked relationship between IQ variability and CA over the age range as a whole, we may accept 16 points as approximately the representa tive value of the standard deviation of IQ s for an unselected popula tion [11] As evidence to justify this position the authors of the scale

[10] From Terman and Merrill *op cit* p 40 (By permission )

[11] *Ibid* pp 39-40 See also M Aborn and G F Derner "IQ Variability in Relation to Age on the Revised Stanford Binet" *Journal of Consulting Psychology* Vol 15 1951 pp 231-235

present the graph shown in Figure 6 3 These distribution curves of composite (L and M) intelligence quotients indicate that their variability is approximately the same for the three age level groupings [12]

Proceeding on the basis of the foregoing reasoning that the standard deviation of IQ s is 16 points and that IQ values are comparable at all age levels, Terman and Merrill provide a table giving intelli-

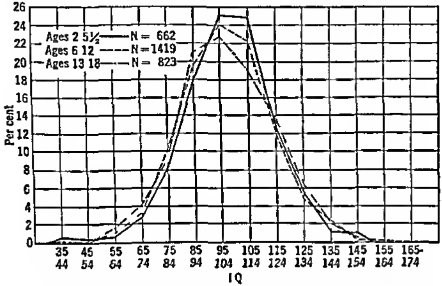



FIG 6 3 *Distribution of Composite L-M IQ s at Three Age Levels* Terman and Merrill *Measuring Intelligence* Boston Houghton Mifflin p 41 (By permission )

gence-quotient equivalents of standard scores (Table 24) This table, if Terman and Merrill s assumption of a normal distribution ' is accepted is useful in giving more precise expression to the *relative* significance of any given intelligence quotient, for by means of an ordinary table of standard deviation values, it is possible to determine the approximate percentile status of any individual in the distribution of intelligence quotients

[12] Some critics have not accepted the argument and conclusions of Terman and Merrill It is our purpose at this stage only to present the scale and the rationale upon which it was constructed by its authors

In respect to the fluctuations in IQ variability, the authors state "Notwithstanding our strenuous efforts to correct for . errors of sampling, complete success is hardly to be expected, and a considerable degree of irregular fluctuation in the found magnitudes of IQ variability from age to age could reasonably be attributed to these sources of error Since inspection of the values reveals no

TABLE 23

*IQ Variability in Relation to Age*[10]

Stanford Binet

| CA | N | S D *Form L* | S D *Form M* |
|---|---|---|---|
| 2 | 102 | 16 7 | 15 5 |
| 2½ | 102 | 20 6 | 20 7 |
| 3 | 99 | 19 0 | 18 7 |
| 3½ | 103 | 17 3 | 16 3 |
| 4 | 105 | 16 9 | 15 6 |
| 4½ | 101 | 16 2 | 15 3 |
| 5 | 109 | 14 2 | 14 1 |
| 5½ | 110 | 14 3 | 14 0 |
| 6 | 203 | 12 5 | 13 2 |
| 7 | 202 | 16 2 | 15 6 |
| 8 | 203 | 15 8 | 15 5 |
| 9 | 204 | 16 4 | 16 7 |
| 10 | 201 | 16 5 | 15 9 |
| 11 | 204 | 18 0 | 17 3 |
| 12 | 202 | 20 0 | 19 5 |
| 13 | 204 | 17 9 | 17 8 |
| 14 | 202 | 16 1 | 16 7 |
| 15 | 107 | 19 0 | 19 3 |
| 16 | 102 | 16 5 | 17 4 |
| 17 | 109 | 14 5 | 14 3 |
| 18 | 101 | 17 2 | 16 6 |

marked relationship between IQ variability and CA over the age range as a whole, we may accept 16 points as approximately the representative value of the standard deviation of IQ s for an unselected population "[11] As evidence to justify this position, the authors of the scale

[10] From Terman and Merrill *op cit* p 40 (By permission )

[11] *Ibid* pp 39–40 See also M Aborn and G F Derner "IQ Variability in Relation to Age on the Revised Stanford Binet," *Journal of Consulting Psychology* Vol 15, 1951, pp 231–235

present the graph shown in Figure 6 3 These distribution curves of composite (L and M) intelligence quotients indicate that their variability is approximately the same for the three age level groupings[12]

Proceeding on the basis of the foregoing reasoning that the standard deviation of IQ s is 16 points and that IQ values are comparable at all age levels Terman and Merrill provide a table giving intelli-

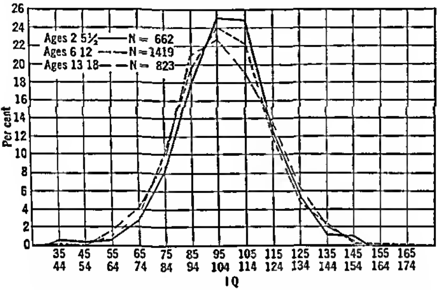



FIG 6 3 Distribution of Composite L-M IQ s at Three Age Levels Terman and Merrill Measuring Intelligence Boston Houghton Mifflin p 41 (By permission )

gence-quotient equivalents of standard scores (Table 24) This table, if Terman and Merrill s assumption of a normal distribution is accepted is useful in giving more precise expression to the *relative* significance of any given intelligence quotient for by means of an ordinary table of standard deviation values it is possible to determine the approximate percentile status of any individual in the distribution of intelligence quotients

[12] Some crit cs have not accepted the argument and conclusions of Terman and Merrill *It is our purpose at th s stage only to present the scale and the* rationale upon which it was constructed by its authors

TABLE 24

*IQ Equivalents of Standard Scores*[13]

*Stanford Binet*

| Standard Score | IQ | Standard Score | IQ |
|---|---|---|---|
| +5.00 | 180 | − .25 | 96 |
| +4.75 | 176 | − .50 | 92 |
| +4.50 | 172 | − .75 | 88 |
| +4.25 | 168 | −1.00 | 84 |
| +4.00 | 164 | −1.25 | 80 |
| +3.75 | 160 | −1.50 | 76 |
| +3.50 | 156 | −1.75 | 72 |
| +3.25 | 152 | −2.00 | 68 |
| +3.00 | 148 | −2.25 | 64 |
| +2.75 | 144 | −2.50 | 60 |
| +2.50 | 140 | −2.75 | 56 |
| +2.25 | 136 | −3.00 | 52 |
| +2.00 | 132 | −3.25 | 48 |
| +1.75 | 128 | −3.50 | 44 |
| +1.50 | 124 | −3.75 | 40 |
| +1.25 | 120 | −4.00 | 36 |
| +1.00 | 116 | −4.25 | 32 |
| + .75 | 112 | −4.50 | 28 |
| + .50 | 108 | −4.75 | 24 |
| + .25 | 104 | −5.00 | 20 |
| .00 | 100 | | |

TABLE 25

*Distribution and Classification of Composite L-M IQ's of the Standardization Group*

| IQ | N | Percent | Classification |
|---|---|---|---|
| 160–169 | 1 | 0.03 | Very superior |
| 150–159 | 6 | 0.2 | |
| 140–149 | 32 | 1.1 | |
| 130–139 | 89 | 3.1 | Superior |
| 120–129 | 239 | 8.2 | |
| 110–119 | 524 | 18.1 | High average |
| 100–109 | 685 | 23.5 | Normal or average |
| 90–99 | 667 | 23.0 | |
| 80–89 | 422 | 14.5 | Low average |
| 70–79 | 164 | 5.6 | Borderline defective |
| 60–69 | 57 | 2.0 | Mentally defective |
| 50–59 | 12 | 0.4 | |
| 40–49 | 6 | 0.2 | |
| 30–39 | 1 | 0.03 | |

[13] From Terman and Merrill *op. cit.* p. 42. (By permission.)

## SUGGESTED CLASSIFICATION OF REVISED STANFORD-BINET IQ'S

The classification in Table 25 has been provided by one of the authors of the 1937 revision[14] It will be noted that the nomenclature and the percents in each of the several categories differ some from those of the 1916 instrument

Like all such tables, its purpose is primarily descriptive and, also, to serve as an aid in the ordering and analysis of testing results The table is valuable, as well, in showing an approximate distribution of intelligence quotients throughout most of the range of mental ability

## ANALYSIS OF FUNCTIONS TESTED

The items of both forms L and M have been analyzed by factorial methods McNemar, who made the first and major analysis, concluded that at each of the several age levels the items test ("are saturated with") a common factor (*g*), and that this common factor is the same one at all age levels (hence it may be called *g*) The *weight* of the common factor differs somewhat among the various age levels, but the common factor accounts, on the average, for about 40 percent of the differences (variance) in scores—hence for about 40 percent of the differences in performance among a group of testees

The statistical results also suggest the presence of group factors at the following ages 2, 2½, 6, 18, and possibly 7 and 11 These are second factors (group factors) that account for from 5 to 11 percent of the differences, while a third factor (another group factor) contributes from 4 to 7 percent The group factors do not appear to be identical at all age levels, nor are they at all well defined as regards the psychological processes involved in them Tentatively, however, McNemar suggests that several of these group factors, at different levels, might be called "memory for designs," "motor," "verbal" *The most definite and significant conclusion, however, is that one factor (g) is sufficient to account for the intercorrelations of test items, with the few exceptions noted*

When complex and comprehensive ("molar" or "global") types of test items are used in a scale, as in the Stanford-Binet (and quite appropriately so), it is not surprising that attempts to isolate group factors through statistical analysis yield results that are indefinite and uncertain, and at best tentative The reason is that these types of items,

[14] From M A Merrill, "The Significance of IQ s on the Revised Stanford Binet Scales," *Journal of Educational Psychology*, Vol 29, 1938, pp 641-651

being complex, involve a number of psychological processes, organized in varying degrees and interacting Group and specific factors will be found most clearly when the test items employed are fractionations and small segments of a whole pattern of mental functioning But such fractionation can destroy "the whole" and can fail to reveal the kinds of mental operations with which the examining psychologist is often most concerned

The identification of a general factor in a revision of the Binet scale should not occasion any surprise, for it will be recalled that Binet himself set out to develop an instrument that should test an individual's general intelligence by means of sampling a variety of mental activities which are manifestations of such intelligence It appears, therefore, that contemporary *statistical analyses, applied to the age* scale, are confirming Binet's psychological insights

It was found, also—as Spearman had shown in his earlier analyses —that the various kinds of test items differed as regards the extent to which they tested (are loaded with) the general factor The following listing shows which items were found to have high loadings of the general factor, and which had low loadings [15]

AGES 2 TO 4½

| *High Loadings* | *Low Loadings* |
|---|---|
| Picture vocabulary | Block building tower |
| Identifying objects by name | Block building bridge |
| *Response to pictures* | *Three hole form board rotated* |
| Comparison balls and sticks | Motor coordination |
| Comprehension | Copying a circle |
| Opposite analogies | Drawing a cross |
| Pictorial identification | Three commissions |
| Naming materials [used in making various objects] | Stringing beads |

AGES 5 TO 11

| *High Loadings* | *Low Loadings* |
|---|---|
| Pictorial likenesses and differences | Paper folding triangle |
| | Patience fitting rectangles |
| Similarities two things | Copying a bead chain |
| Vocabulary | Copying a bead chain from memory |
| Verbal absurdities | |
| Similarities and differences | Picture absurdities |
| Naming the days of the week | Word naming [free association] |
| Dissected sentences | Word naming animals |
| Abstract words [definitions] | Block counting |

[15] From McNemar, *op cit* pp 111–113

AGES 12 TO SUPERIOR ADULT III

| *High Loadings* | *Low Loadings* |
|---|---|
| Vocabulary | Problems of fact |
| Verbal absurdities | Copying a bead chain from memory |
| Abstract words [definitions] | Memory for stories |
| Differences between abstract words | Enclosed box problem |
| Arithmetical reasoning | Paper cutting [visual imagery] |
| Proverbs | Plan of search |
| Essential differences | Repeating digits [forward] |
| Sentence building | Repeating digits reversed |

Examination of the items that have low loadings reveals that they test only a very limited range of functioning and that, with two exceptions, they involve only the following processes visualization (space perception and spatial relationships), visual imagery, and rote memory (immediate recall) All of the test items, among the low loadings, falling under the foregoing categories are lacking, relatively, in complexity and would, therefore, not have the differentiating power of the more complex tasks required by the items under high loadings The exceptions are word naming (random and animals), and problems of fact The latter is properly considered to be a test of reasoning Its low loading with *g* is therefore surprising, but no explanation is apparent or available Random word naming tests richness of free association, word naming of animals tests controlled association A possible explanation of their low loadings might be that they are fairly routine tasks that do not require the reasoning (organization, analysis) demanded by the items having high loadings within the same range of ages (5–11)

The foregoing listings are significant for at least two additional and important reasons First, a knowledge of test items that have high or low loadings of the general factor enables the examining psychologist to make a more thorough analytical and meaningful evaluation of an individual s over-all test performance The examiner is thus in a better position to evaluate the strength of the general factor in a particular testee This is particularly valuable if the psychological nature of the general factor has been determined Second, inspection of the lists of items having high loading strongly indicates that the general, or common, factor is one that involves acquisition of, use of and reasoning with symbols—namely, language and number—even though the testing of these begins at a very elementary level and at times utilizes nonverbal materials in presenting the problem (e g , pictures, sticks).

The mental activities required by these test items have very much in common with Spearman's view that intelligence is essentially the ability to educe relations and correlates

Specifically, the following processes are involved in the test items having high loadings acquisition and use of vocabulary, verbal analysis of a situation, verbal and numerical concept formation, insights into similarities and differences (also involving concept formation); analysis and synthesis of materials, both nonverbal and verbal, organization and reorganization of materials, both nonverbal and verbal

The list of items given above (based on statistical calculations) and the indicated psychological functions that are involved provide an illustration of how statistical and psychological analyses work together They also make it clear that superficial observations of differences between test items can be misleading as to their essential psychological processes For example, the test items at early age levels requiring identification of objects by name or use might be regarded simply as tests of information or of specific rote learning, whereas they actually have much in common with items that test "comprehension," and which are more obviously tests of reasoning Or, at a somewhat later age level, ability to define certain words (vocabulary test) might be regarded simply as the result of specific learning and verbal facility, whereas actually it has much in common with perception of pictorial (as well as verbal) similarities and differences

It is useful to classify test items as "information," "word knowledge," "perception of forms," "reasoning," etc , but the point is that such classification does not necessarily signify that each of the subtest classifications measures a distinct group-factor or a special factor [16]

---

[16] McNemar's findings are in close agreement with those of an independent study made in Great Britain Cyril Burt reports that a common factor accounts for 42 percent of Stanford Binet test variance, and that two subsidiary factors account, respectively for 12 percent and 16 percent of test-score differences The close correspondence obtained in the United States and in England gives additional weight to and confidence in the Stanford Binet Scale as an instrument for measuring intelligence, particularly of children and adolescents See Cyril Burt and Enid John "A Factorial Analysis of Terman Binet Tests Part I, *British Journal of Educational Psychology*, Vol 12, 1942, pp 117-127, *ibid*, Part II, Vol 12, 1942, pp 156-161

For a dissenting analysis and interpretation see L V Jones "A Factor-Analysis of the Stanford Binet at Four Age Levels," *Psychometrika* Vol 14, 1949, pp 299-331

## TYPES OF ITEMS

Bearing this important distinction in mind then we may indicate the types of items included in the Stanford Binet Scale [17]

| *Test Items* | *Functions Involved* |
|---|---|
| | *Years 2–5* |
| Form perception and manipulation (blocks form boards stringing wooden beads)<br>Perception of differences in size and form | Visual perception and analysis |
| Visual motor operations | Visual analysis plus motor development |
| Perception of relationships (in pictures) | Visual perception plus beginnings of concept formation |
| Rote memory (using digits and sentences) | Immediate recall |
| Use of words in combination<br>Identifying objects by name or use<br>Following directions | Language development and comprehension |
| Verbal comprehension and word knowledge<br>Understanding of opposites | Reasoning with abstractions and concept formation |
| | *Years 6–12* |
| Form perception | Visual analysis |
| Visual motor operations | Visual analysis plus motor development |
| Rote memory (using digits and sentences) | Immediate recall |
| Word knowledge (concrete and abstract) | Language development |

[17] Since the Bellevue Scale utilizes many of the same types of items as the Stanford Binet additional factors affecting test performance on these items will be presented in connection with the Bellevue Scale in the next chapter

| *Test Items* | *Functions Involved* |
|---|---|
| Verbal comprehension | Reasoning with abstractions and concept formation |
| Number concepts<br>Arithmetical reasoning | Number concept formation and reasoning with abstractions |
| | |
| ***Year 13—Superior Adult III*** | |
| Visual analysis and imagery<br>Perception of visual relationships<br>Visual motor operations | Visual perception and analysis plus reasoning with non verbal materials |
| Rote memory (using digits words and sentences) | Immediate recall |
| Word knowledge | Language development |
| Synthesis of verbal materials<br>Problem solving, using verbal materials | Reasoning with abstractions |
| Verbal analysis<br>Arithmetical problems<br>Analysis and comprehension of symbols | Concept formation plus reasoning with abstractions |

## THE SHORT SCALE

It is possible to administer an abbreviated form of the scale the constituent items having been specified by the authors A short scale presumably is used when the examiner does not need as accurate an index of measurement as it is possible to obtain and when the necessary time is not available for a full length examination The use of an abbreviated form however should be discouraged for when it is at all desirable to administer a test of mental ability, it would be very unwise indeed not to require the greatest possible accuracy [15]

---

[15] For some results obtained with the short scale see for example G Spache "Methods of Predicting Results of Full Scale Stanford Binet," *American Journal of Orthopsychiatry* Vol 14 1944 pp 480–487 G Spache "The Abbreviated Stanford Binet Scale in a Superior Population" *Journal of Educational Psychology* Vol. 35 1944 pp 314–318

## EVALUATIONS AND CRITICISMS

Every scale, quite properly, is subjected to evaluation and criticism on both theoretical, practical, and experimental grounds The 1937 Stanford Binet is no exception On the whole, educators and clinicians who have had experience with both the old and the new Stanford-Binet scales are in essential agreement that the new one, in spite of some inadequacies, is a more useful instrument of its kind than the older one The following, however, are some of the questions and criticisms that bave been raised

**Is the age-scale (Binet) type of test preferable to the point-scale type of test?** The standardization of an age scale is much more laborious and rigid than standardization is in the case of a point scale The result is that even when experience and experimentation reveal certain defects and inadequacies in an age scale, the difficulties in making the indicated changes are so great as to be a deterrent to early revision

As instances in point, some psychologists report difficulties in interpreting responses to certain items and in scoring them They also question the age placement of some items The task of correcting these defects, if the criticisms are warranted, would be great [19] Note, for example, that the authors of the 1937 Stanford Binet made six revisions of Form L before they were sufficiently satisfied with the age placement and grouping of test items, which would yield correct mental ages, intelligence quotients, and IQ distributions

In the case of a point scale, on the other hand, it is a much simpler process to revise age norms All other things being equal, of course, the simpler and easier methods should be employed to achieve a desired goal in psychological testing But simplicity and ease alone should not be the decisive considerations The crucial question is whether the age-scale type of test or the point scale type provides a superior means of obtaining a measure of an individual's mental ability Thus far, although the views of competent professional persons are not unanimous, it appears that the age scale is preferable for use with children and young adolescents But when older adolescents

[19] For example M Krugman, "Some Impressions of the Revised Stanford Binet Scale *Journal of Educational Psychology*, Vol 30 1939, pp 594–603, H E Garrett, "The Standardization of the Terman Merrill Revision of the Stanford Binet Scale," *Psychological Bulletin* Vol 40, 1943, pp 194–201

and adults are to be examined, it has often been found that point scales (e g , the Wechsler Bellevue) are preferred [20] Also, there is at present a trend toward increased use of point scales because of their apparent value in clinical diagnosis by means of "scatter analysis" of scores on the subtests. (See Chapter 15 )

**Do the differences in variability (standard deviation) at the different age levels seriously impair the usefulness of the Revised Stanford-Binet Scale?** A wholly satisfactory test of intelligence should have equal or very nearly equal variability at all age levels [21] Data already

TABLE 26

*IQ's Adjusted for Variability Differences at Several Age Levels* [22]

| | *Adjusted IQ s* | | | |
|---|---|---|---|---|
| *Obtained IQ s* | *2–4 to 3–3* | *4–10 to 6–6* | *11–6 to 12–5* | *14–6 to 15–5* |
| 140 | 134 | 149 | 135 | 136 |
| 130 | 125 | 137 | 126 | 127 |
| 120 | 117 | 125 | 117 | 118 |
| 110 | 108 | 112 | 109 | 109 |
| 100 | 100 | 100 | 100 | 100 |
| 90 | 92 | 88 | 91 | 91 |
| 80 | 83 | 75 | 83 | 82 |
| 70 | 75 | 63 | 74 | 73 |
| 60 | 66 | 51 | 65 | 64 |
| 50 | 58 | 39 | 56 | 56 |

cited (Table 23) show that the standard deviations of IQ's as found by Terman and Merrill were not equal at all age levels, though most fell close to 15–17 points However, data have been provided to adjust for variability differences in IQ at age levels where such adjustments are necessary Table 26 provides some examples of these adjustments

Table 26 is read thus If an individual's age is between the limits of 2–4 and 3–3 and his obtained IQ is 140, his adjusted rating would be 134

Two aspects of this table should be noted in particular First, the

[20] For example F Halpern, "A Comparison of the Revised Stanford L and the Bellevue Adult Intelligence Test as Clinical Instruments," *Psychiatric Quarterly Supplement* Vol. 16 1942, pp 206–211

[21] This requisite is based upon the principle that the distribution of general ability is equal at all the age levels covered by the test, rather than being either irregular from year to year or changing systematically at successive ages

[22] From Q McNemar, *op cit*, pp 173–174 (By permission )

adjustments are small for IQ's near the average (100) and larger as the obtained IQ's deviate more from the average Second, in no instance would the adjusted IQ seriously displace an individual from one *general* level of ability to another, even though in a few instances the changes are appreciable, notably from an IQ of 50 to one of 39 in the 4–10 to 6–6 range

The answer to our question is, then, that differences in obtained IQ variability on this scale do not seriously impair the usefulness of the Revised Stanford Binet Scale, and they *need not* impair its usefulness if the user of the scale is aware of the differences in variability at the several age levels and makes the necessary adjustments of IQ values It is to be expected, of course, that later revisions of the Stanford-Binet will minimize the adjustments [23]

**Are the Stanford-Binet test items a variety of disconnected tests?** This scale has been criticized at times as being only that This criticism, however, is based upon a failure to take into account the theory of intelligence and the method of measurement upon which the scale is based namely, the sampling of *general ability* by means of a wide variety of types of items in order to obtain an adequate estimate of the processes involved Furthermore, the factorial analyses already discussed in this chapter show that the test items measure primarily a general factor that is common to all age levels of the scale [24] On cursory inspection, the items may *seem* to be dissociated, but psychological and statistical analyses demonstrate that such is not the case in fact

**Is the composite, or "global," type of scale (such as the Stanford-Binet) preferable to the factorial analyzed type, in which "factors" or "primary abilities" are separately tested and scored?** Again, the final answer to this question will depend upon whether the factorial type

[23] The cause of the differences in standard deviations in IQ at different age levels may be due to either or both of the following (1) unequal item difficulty at the several age levels (2) inadequate samplings of the standardization population See McNemar, *op cit* Chapter 8 also A L Baldwin, "Variation in Stanford Binet IQ Resulting from an Artifact of the Test,' *Journal of Personality*, Vol 17, 1948, pp 186–198 J A F Roberts and M A Mellone, "On the Adjustment of Terman Merrill IQ s to Secure Comparability at Different Ages," *British Journal of Psychology, Statistical Section 5*, 1952, pp 65–79

[24] Even the analyses that emphasize group factors rather than a general factor do not support the criticism that the items are merely "a variety of disconnected tests"

of scale proves to be more valid, more accurate, and more useful than the composite type has been in clinical work and in educational and vocational guidance

It is not improbable, however, that psychologists will develop satisfactory tests of intelligence that will yield an index of general ability (e g , mental age and intelligence quotient) and which, at the same time, can be analyzed and scored for particular aspects of intelligence or particular abilities To do this, instead of grouping items by age level, it will be necessary to devise a number of parts, or subtests (e g , verbal reasoning, numerical ability, etc ), each of which would contain a series of items, all items within each subtest measuring the same mental processes, but scaled in difficulty Each subtest would yield a separate index of test performance, and at the same time the individual's average performance would represent his general level This is what the Wechsler scales (discussed in the following chapter) aim at

At the present time, the Binet type of test provides only the general indexes of MA and IQ But—and this matter will be dealt with more fully later—the qualified examiner is able to make a significant *qualitative* analysis of a subject's performance on the scale

**Is the Stanford-Binet Scale too heavily weighted with verbal materials?** A criticism, heard with some frequency against both the old and the new revisions, is that these scales place a premium upon "verbal intelligence" and that subjects having language handicaps are penalized and incorrectly rated In reply to this criticism, Terman and others hold that the most essential and most significant aspect of higher thought processes is the ability to do conceptual and abstract thinking, that is, to operate with language, number, and other symbols It is maintained, also, that the vocabulary test, when used with children from homes where English is the primary language, has higher value than any other part of the scale

It must be emphasized, however, that this is not so in the case of a child who, even though he comes from such a home, has reading or language difficulties due not to lack of capacity but to visual or auditory defects or anomalies

In actual clinical practice, the examiner should always supplement an essentially verbal test of intelligence with one of the nonverbal type if he has any reason to suspect that the former penalizes the subject There will be occasions also when it will be desirable to obtain a

rating for an individual on both types even where no language handicap is indicated for the purpose of comparing two or more aspects of a subject s ability While the correlation between performance on verbal and on some nonverbal tests of mental ability is high coefficients of correlation reflect *group* trends and relationships and unless the correlation coefficient is perfect (plus or minus 1 00), there are always individual exceptions from the generalization that can be made on the basis of the coefficient, hence the need at times for the study of the several aspects of ability in an individual case

**Is the Stanford-Binet Scale a test of school learning?** As stated in an earlier chapter any scale must test mental ability through its manifestations through activity of some kind And as Binet originally pointed out the tests must be adapted to the environment of the subjects to be rated It is reasonable and sound therefore that a scale designed to test mental ability primarily of American children and adolescents should utilize the *effects* of common schooling experiences as well as effects of some other common experiences To say that the Stanford Binet or any other scale is only a measure of school learning is unwarranted Differences in quality and extent of opportunity to learn in school and out, will have an effect upon intelligence-test scores, but such differences in opportunity do not in themselves account for all individual differences in ability that are found Good schooling and other good environmental conditions nurture an individual s mental capacities and provide optimal conditions for his mental development Thus we can say that the Stanford Binet and other scales provide ratings of intelligence within limits of error, under existing school conditions, general environmental conditions, and clinical conditions Obviously, therefore the examiner must know the general developmental background of the individual he is testing if his interpretation of test results is to have real validity

Several studies of satisfactory and unsatisfactory responses on the Stanford Binet have found that bright and superior children answer correctly more of the intellectual items than do normal or dull children [25] These items include the verbal and numerical utilizing symbols

[25] See A L Baldwin "The Relative Difficulty of Stanford Binet Items and Their Relat on to IQ *Journal of Personal ty* Vol 16 1948 pp 417–430 A Magaret and C W Thompson Differential Test Responses of Normal Superior and Mentally Defective Subjects" *Journal of Abnormal and Social Psychology* Vol 45 1950 pp 163–167

and abstractions This finding does not mean that bright and superior children have attained their ratings on mental tests only as a result of schooling It is to be expected that individuals who are potentially above average in mental ability will be superior in dealing with situations and problems employing language and number, for the greater the capacity of the individual is for mental development, the greater will be his ability to deal with symbols and to handle situations and problems at the level of abstraction The converse has also long been known, namely, that one of the principal deficiencies of the mentally retarded and mentally defective is their inability to deal with materials and concepts at the levels of abstraction It will be recalled that one definition states that intelligence is the ability to deal with abstractions It will be recalled, also, that the educing of relations and correlates extends upward to the use of symbols (language and number)

**Are some of the items in the Stanford-Binet Scale obsolete?** Since any test utilizes materials from the environment in which it is to be used, and since environments normally undergo change, it is to be expected that some items in any given test will in time become culturally obsolete In the Revised Stanford-Binet there are a few such items For example

Identifying a toy steam locomotive by name
Identifying objects (pictures) by use, such as an old fashioned kitchen stove
*Response to a picture (Messenger Boy, year 12)*

There are not many such items in this scale In the case of some of the non-obsolete items, however, it becomes necessary, with time, to revise to some extent the responses that are acceptable for credit For example, the following item is one such "What's the thing for you to do when you are on your way to school and see that you are in danger of being late?" (Year 7) When scoring an unusual response to an item like this, the qualified examiner is warranted in exercising his judgment as to its correctness, in fact, he has to do that And his decision will be based upon his familiarity with the psychological processes being tested by the particular item

**Does the Stanford-Binet Scale test different abilities at different age levels?** The answer to this question is to be found largely in the preceding discussion of "Analysis of Functions Tested" in this chapter.

There it was stated that the two major analyses (by McNemar and Burt) found that the scale measures principally a general factor that is common to all age levels, and in addition there appear to be group factors at 2, 2½, 6, 18, and possibly at 7 and 11 It was pointed out that the items baving low loadings of the general factor were, with minor exceptions, tests that required the use of visual perception, visual imagery, and rote memory Emphasis was placed, also, upon the necessity of distinguishing between item-type classifications of tests (vocabulary, arithmetical reasoning, etc ) and basic psychological processes involved in each

**Does the scale measure originality and creative abilities?** The answer is that it does not measure these abilities, as such, to an important degree This aspect of intelligence was discussed in Chapter 3, "Definitions and Analyses of Intelligence " There it was pointed out that the requirement of *objectivity* in standardizing and scoring tests practically excludes tests of creativeness and originality But, as also stated, while not all individuals who achieve high scores on intelligence tests evidence originality and creative ability, those who are capable of originality and creative mental activity do generally obtain high test scores It is possible to say, therefore, that persons having creativeness and originality will be found generally in the group who attain superior ratings on the intelligence scales Furthermore, although originality and creativeness cannot be included in the prescribed objective scoring, a qualified examiner will note responses indicating these traits and will include them in his interpretation and evaluation of the examinee's performance

**Is the 1937 Stanford-Binet Scale adequate at the adult level?** The standardization group of this scale included individuals of 18 years, but it did not include an adult population Therefore, the test items at the several adult levels rest upon theoretical considerations already mentioned, rather than upon actual samplings of adult performance One result, due perhaps to the methods used in standardizing the scale at superior adult levels, has been its frequently observed inadequacy with college students who, as a group, would be ranked above average The inadequacy of the scale is especially marked when administered to very superior students, for it is not difficult enough at the higher levels of adulthood

As new scales are devised—built upon different conceptions and theories—they will have to be subjected to both experimental investigation and practical use before their value can be compared with that of the Stanford Binet A valid judgment cannot be reached on *a priori* grounds In the meantime, the great value of the Stanford Binet Scale having been demonstrated, psychologists will continue to use it They will bring to bear their insights on the interpretation of the *behavior* of the persons being examined and on the interpretation of the *test results* obtained [28]

---

[28] The reader will find a very useful analytical table giving the content of the several Binet scales and of the several American revisions in G D Stoddard, *The Meaning of Intelligence* New York Macmillan, 1943, pp 483–497

In a personal communication [March, 1954] Dr Maude A Merrill wrote "Analysis of my preliminary sample of some 800 cases [tested within the last five years] between five and fourteen yields results that are very heartening The indications are that the scale is doing in 1954 pretty much what it did in 1937, that there are some differences in difficulty in certain of the subtests and specifically certain of the items that can be changed in accordance with the indications by adding pretested substitute or modified items without doing violence to the original structure of the scale The curves showing increase in percent passing with age are remarkably regular and follow very closely the original percentages except on a rather surprisingly small number of items One finding was that the modern elementary school child is handicapped on tests that assume the acquisition of reading skills In 1937 we could assume that, on the average, by the age of ten, children had acquired reading techniques that would enable them to make use of such skills in the problem solving tasks presented by certain of the subtests *Reading and Memory* at the ten year level, *Minkus Completion* at the twelve year level and *Dissected Sentences* at thirteen stand out as presenting much more difficulty than in 1937 The obvious inference is that this finding reflects differences in educational practices These will need to be verified on another sample" (Reproduced with permission of Dr Merrill)

# 7.

## THE WECHSLER SCALES

THE Binet scale and its several revisions are largely verbal in content, although some nonverbal items are included, especially at the early age levels There are, however, other scales which are wholly, or in large part, of the performance or nonverbal type This type of test is one in which the use of language is eliminated from test content and response, although directions are generally given orally In a few instances the directions, too, are given without the use of language, by employing pantomime instead.

The test materials of a nonverbal scale consist of concrete objects such as form boards, cubes (to be arranged in specified ways), mazes, geometric figures, pictures (cut up, to be correctly assembled), and others that will be described in later sections The individual's responses take the form of manipulations, visual perceptions, and interpretations which are implied by what he *does* rather than by anything he says

Performance tests were first devised as a supplement to or substitute for the Stanford Binet scale in order to examine deaf, illiterate, or non English speaking subjects Since their introduction, the use of nonverbal tests has been extended, for they are now utilized with children who have or are suspected of having reading difficulties, with those who have attended school irregularly and might thus have been handicapped in developing verbal ability, and with persons who might have been handicapped by markedly inferior environmental conditions Nonverbal tests are used, also, by examiners who, for any other reason, believe that such a scale will yield a more complete picture

of the individual whose capacities are being analyzed and evaluated

The Wechsler scales, to be described in the following pages, combine verbal and nonverbal materials within a single instrument in an effort to obtain the advantages, comparisons, and contrasts yielded by both types of materials

## DESCRIPTION OF THE WECHSLER-BELLEVUE INTELLIGENCE TEST[1]

This scale, published in 1939, is intended to test the intelligence of persons from the age of 10 through the age of 60 years, although norms are available beginning at 7 5 years This or a similar beginning level is necessary, of course, if adults and adolescents of lower-than-average mental levels are to be tested by means of the scale

The Bellevue scale will be presented in some detail so that the reader may understand the rationale of its construction, its values, and its limitations

**Content of the Scale.** The scale is in part verbal and in part performance, enabling the examiner to obtain three scores and three quotients a full-score and its intelligence quotient, a verbal score and its IQ, a performance score and its IQ

The scale has been constructed in this form on the principle that intelligence involves not only ability to deal with symbols, abstractions, and conceptual thinking, but that it also involves ability to deal with situations and problems in which concrete objects, rather than words and numbers, are utilized The scale also is put to the pragmatic test of whether a given combination of items (in this instance verbal and nonverbal) serves the purpose of individual mental examination and analysis of capacities better than other combinations The types of tests included in this scale are not unique They were selected from available sources after a study had been made of a variety of standardized tests then in use The objective was the construction of an

---

[1] Cf D Wechsler, *The Measurement of Adult Intelligence* Baltimore Williams and Wilkins 1944 third edition A second form of this scale was published in 1946 Form II is identical with the first form in respect to underlying principles and types of test materials The specific content of course is different See *The Wechsler Bellevue Intelligence Scale Form II* New York The Psychological Corporation 1946

effective scale for adolescents and adults, based upon already known and proven psychological materials and procedures

The Bellevue scale consists of eleven parts, or subtests, six being verbal in content and five nonverbal[2] The principal difference between the Stanford-Binet and the Bellevue scale, in respect to arrangement of items, is this in the former, items of various types, testing a

FIG 7 1 *A 12 year-old girl is shown here taking the block-design test of the Bellevue scale The* model of the design to be copied is on the card before her *The examiner is timing her with the stop-watch in his left hand* (Acme Photo )

variety of functions, are grouped together at each age level, in the latter, all items of one type are grouped together, constituting a subtest of the whole In the case of the latter, the effort is made to arrange *the individual items within each subtest in a sequence of increasing* difficulty A description of the subtests follows The first six constitute the "verbal scale", the remaining five constitute the "performance scale"

[2] One of the verbal subtests vocabulary, was originally added as an alternate or supplement. However it is now used quite regularly with the other subtests.

(1) *Information* This consists of twenty five items of information, covering a wide range (For example, "How many weeks are there in a year?") The assumptions are that the questions cover a wide enough range of materials to provide an adequate sampling of information acquired by a person who has had the usual opportunities of our society, that the range of an individual's information is an indication of his intellectual capacity, and that the more intelligent have broader interests, more curiosity, and seek more mental stimulation *This view can be valid, however, only if the subjects being tested have had the usual opportunities for experience and learning and if the test items are a valid sampling of the opportunities to acquire information* It is necessary to keep this caution in mind in evaluating a test of information Performance on information tests is susceptible, also, to variations in individual motivation, i e, some bright, self absorbed persons or those unreceptive, for emotional reasons, to the offerings of their environments have unduly limited funds of information Others, for motivational reasons quite removed from general intelligence, exhibit an exaggerated and misleading fund of information For these very reasons, however, a test of information is a useful one for clinical purposes that is, for purposes of diagnosing personality traits

(2) *General comprehension* This part of the scale consists of ten problem situations (plus two alternates), in which the subject must comprehend what is involved in the situations and provide answers to problems presented (For example "Why should people pay taxes?") Success on this subtest, it is held, depends upon possession of practical information plus ability to evaluate and utilize past experience It appears, also, that ability to verbalize is a factor contributing to success Tests of general comprehension are now very commonly used in intelligence scales, and, it will be recalled, they were included in the Binet scales They have been found valuable clinically in revealing the thought processes, background, feelings, and emotions of the subject

(3) *Arithmetical reasoning* This part of the Bellevue scale is designed to test "mental alertness" The problems, the author of the test states, do not require "knowledge" (presumably arithmetical *skills*) beyond that of the seventh grade level Problems in arithmetical reasoning are widely used in tests of intelligence, since they have been found to have significant correlation with total scores of scales and to have high predictive value in respect to future evidences of mental ability

(4) *Memory span for digits forward and backward* The subject is required to repeat series of digits heard once The series vary in length from three to eight (backward) and nine (forward) This is a test of immediate recall, or immediate memory span Psychological studies—both experimental and clinical—have consistently shown that tests of memory span, or immediate recall, of digits have a low correlation with other, more valid tests of intelligence Yet, memory span for digits continues to be used because it is helpful in detecting the mentally defective, whose span is often very short (generally less than five digits forward and less than three backwards), and because very poor span is useful in making certain clinical diagnoses of organic defects Poor memory span for digits, especially backwards is also found at times in cases of persons who are unable to apply the attention necessary in solving more difficult mental tasks

(5) *Similarities* This part of the scale consists of twelve sets of paired words, for each pair the subject is required to state the similarity or similarities that exist (For example, orange banana ) The author of the scale regards the similarities test as one of the most satisfactory, for it appears to sample very well the ' general factor (Spearman's *g*), or what is commonly called *general intelligence*

(6) *Vocabulary* This test consists of forty two words selected from an original list of one hundred which were chosen from a dictionary according to a sampling formula [3] and then experimentally arranged in order of difficulty The reader is already familiar with the view held by many psychologists that a vocabulary test—where there have been no unusual developmental factors—is one of the most valuable kinds of materials in deriving an index of a person s general intellectual ability Thus although the vocabulary list was originally recommended as an alternate test in the Bellevue scale, experience demonstrated its value, so that it is now suggested that this part be included regularly when the full scale is to be administered Also, like Binet and many other psychologists, users of the Bellevue scale observed that qualitative differences in word definitions, given by various subjects, have clinical value in helping to reveal the nature of an in

[3] "The words were taken from one of the Funk and Wagnall s Standard (School) Dictionaries The list was arrived at by choosing 100 words at random in the following manner Beginning with an odd page we selected every top word but one (omitting however obsolete technical or esoteric words) in the left hand column of every fifth page and continued the process until we had 100 words" (Wechsler *op cit* p 99 )

dividual s thought processes (their depth, extent of analysis, nuances of meanings, queerness of definitions, cultural background) and, in some instances, feelings, emotions, and values

(Numbers 6 through 10 are the performance tests )

(7) *Picture arrangement* In this subtest of the Bellevue scale, there are seven series of pictures Each series is presented to the subject in a disarranged order, but when the pictures in each series are placed in the correct sequence, they tell a story This type of test, it is held, measures a person's ability to comprehend and evaluate a total situation without the use of language From the author's data, however, it appears to involve the 'general factor' only to a moderate degree, but sufficiently to make a contribution to the total sampling

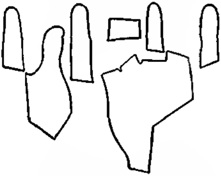


FIG 7 2 *The Disassembled Hand (Object Assembly Test)* From the Wechsler Bellevue scale (By permission )

(8) *Picture completion* In this part, there are fifteen cards, each of which shows a picture that is incomplete in some detail (for example, a picture of a face with the nose missing) The testee is required to note and name the missing part [4] In some pictures the task is quite simple for the ordinary person, but in others the deficiencies of the pictures are somewhat more subtle It has been found that this material is particularly valuable in testing lower level intelligence, as well

[4] Generally this type of test is called mutilated pictures " It was used by Binet in his scales and is now widely used in group tests as well as in Binet revisions The term *Picture Completion* is misleading since the testee does not actually complete the picture

as having moderate discriminative value at the intermediate levels At *the higher levels, however, this test is inadequate because it is not of* sufficient difficulty On the whole, it is said that this part of the scale " . measures the individual s basic perceptual and conceptual abilities in so far as these are involved in the visual recognition and identification of familiar objects and forms . . In a broad way, the test measures the ability of the individual to differentiate essential from nonessential details "[5]

(9) *Block design* This test utilizes sixteen identical cubes, some or all of which are used to make nine given designs (two of which are for demonstration) One side of each cube is colored blue, one red one white, one yellow, one red and white, one blue and white (the last two being divided diagonally) For each of the four easiest designs, only four blocks are used, for the next two, nine blocks are used, for the most difficult design, sixteen are used The author of the scale believes that this test involves ability to analyze and synthesize He reports that scores on block design correlate well with total scores of the test, as they do also with the separate scores on the comprehension, information, and vocabulary tests of the Bellevue scale This would indicate that the block-design test is valuable as a measure of the general factor

(10) *Object assembly* This test includes three "figure formboards", that is, three familiar objects made of wood, each one cut into several parts which have to be assembled to make the whole The objects are a *manikin* a *feature profile* (side view of a human head), and a *hand* The inclusion of this test is justified by the author on the ground that it is desirable to have a device which requires the subject to perceive and reconstruct the parts of familiar objects into their wholes Evidence on the object assembly tests indicates that it correlates poorly with other parts of the scale and has only very limited value in differentiating between individuals It does, however, have clinical value, of a qualitative kind, in that it contributes to the examiner s understanding of the subject s 'modes of perception, the degree to which one relies on trial and error methods, and the manner in which one reacts to mistakes'[6] This subtest is useful, also, in diagnosing disturbance of visual perception

[5] Wechsler *op cit* pp 90-91
[6] Wechsler, *op cit.*, p 93

(11) *Digit symbol test* The subject is shown nine divided rectangles, in the upper half of each rectangle is a digit, in the lower half there is a symbol For example, the following

| 1 |
|---|
| — |

The key is followed by 75 rectangles (of which eight are practice samples) in which only the numerals are given In each instance, the subject is required to add the appropriate symbol This test, also known as a substitution test, is regarded as requiring the association of symbols and involving speed and accuracy of performance It involves, also visual memory The purely motor factor it has been found is relatively unimportant except in the case of illiterate persons who are not accustomed to using pencil and paper

## FUNCTIONS INVOLVED IN THE SUBTESTS

The functions involved in each of the eleven subtests may be psychologically analyzed as shown below They indicate the processes that are operative in most effective performances on each of the tests This analysis should be distinguished from a *factorial* analysis, which is a *statistical* procedure whereby an attempt is made to consolidate and reduce the number of nonstatistically analyzed functions on the basis of communality, and perhaps to re name them

| *Subtest* | *Functions* | *Influencing Factors* |
|---|---|---|
| Information | Long range retention<br>Association and organization of experience | Cultural environment<br>Interests |
| Comprehension | Reasoning with abstractions[7]<br>Organization of knowledge | Cultural opportunities<br>Response to reality situations |

[7] "Reasoning with abstractions" generally involves the processes of both analysis and synthesis with the use of symbols—language and number The testee must first analyze the relationships existing among the members or parts of the whole problem then he must reorganize and interpret and at times create new wholes in order to reach the desired solution

| *Subtest* | *Functions* | *Influencing Factors* |
|---|---|---|
| Arithmetic | Reasoning with abstractions<br>Concept formation<br>Retention (of arithmetical processes) | Attention span<br>Opportunity to acquire the fundamental arithmetical processes |
| Similarities | Analysis of relationships<br>Verbal concept formation | A minimum of cultural opportunities |
| Vocabulary | Language development | Cultural opportunities |
| Digit Span | Immediate recall<br>Auditory imagery<br>Visual imagery at times | Attention span |
| Picture Arrangement | Visual perception of relationships (visual insight)<br>Synthesis of nonverbal material | A minimum of cultural opportunity<br>Visual acuity at times |
| Picture Completion | Visual perception analysis<br>Visual imagery | Environmental experience<br>Visual acuity at times |
| Object Assembly | Visual perception synthesis<br>Visual motor integration | Rate of motor activity<br>Precision of motor activity |
| Block Design | Perception of form<br>Visual perception analysis<br>Visual motor integration | Rate of motor activity<br>Minimum of color vision |
| Digit Symbol | Immediate rote recall<br>Visual motor integration<br>Visual imagery | Rate of motor activity |

While the ability to verbalize and to make abstractions is not necessarily involved in the five nonverbal subtests, nevertheless it has often been observed by examiners that this ability does facilitate and expedite one's performance For example, on the Block-Design test, it is possible to analyze and formulate the color and form relationships

of each design before beginning to reproduce it In the Picture Arrangement test, some subjects will attempt to discern and state the story told by the group of pictures before starting to place them in the correct sequence It is important to recognize this psychological fact in evaluating and analyzing performance on tests that are primarily nonverbal namely, *that even with such types of tests, ability to verbalize and abstract may be and at times is one of the psychological functions involved*

## NEED FOR AN ADULT SCALE

It will be recalled that Binet's own scales were not suited to use with adults, nor was the Stanford Revision of 1916 And while the 1937 Stanford-Binet is better standardized at the upper levels, including three superior adult" levels, there is reason to question its adequacy in testing superior adults There is also, as has already been explained, the difficulty and inconsistency of using the mental-age index with superior adults, that is, those having intelligence quotients above 100, who would therefore have to have a mental age rating above that of the average adult

Intelligence testing of adults was begun on a large scale in 1917 with the establishment of a psychological division in the United States Army, in World War I At that time, the Army Alpha (verbal) and Army Beta (nonverbal) tests were assembled, and with them about one and three-quarters million men were tested This experience in large-scale testing provided the impetus for the development, after the war, of a number of other group tests for adults But these tests did not prove to be adequate, with the exception of several which were designed for use with selected and limited groups of our population, such as candidates for admission to colleges (These will be presented in a later chapter ) The Bellevue scale was, therefore, developed and has been offered as an *individual* test, standardized for ages ten to sixty[8]

## STANDARDIZATION

The first step in standardizing the Bellevue scale was to adopt a view of intelligence which should serve as the framework, so to

[8] Although the age distribution of subjects used in standardizing the scale was from seven to seventy, the author gives the age range for use as ten to sixty

speak, within which the test items would have to fit The general factor ($g$) theory was adopted, which requires that there should be significant intercorrelations between the several parts of the scale, and that the scale in its totality should provide a valid index of the individual s general ability

Having accepted the theory of a general factor, the author of the Bellevue scale then had to determine which types of test materials should be included in the scale in order best to measure that factor On the basis of past experience, it was decided that both verbal and nonverbal materials provide the most adequate and representative content, rather than either one alone Having arrived at this view, the author and his collaborators proceeded to select the particular *types* of subtests which had been widely used by many other psychologists and which experience and experimentation had proved were valuable In some instances, specific *items* already in use, were included within appropriate subtests, in other instances, it was necessary to create new items for each of the subtests

### THE POPULATION SAMPLE

The next step was the selection of the population upon whom the scale should be standardized The basis used was this the sampling of adult population should be based upon the occupational distribution of the country s adults, as shown in the United States Census of 1930 The very numerous occupational subdivisions of the census were combined to make ten comprehensive categories (such as agriculture, manufacturing clerical etc ) From the records of many adults who had already been tested and whose scores were available, a standardization population was selected so that their occupational distribution should correspond reasonably well with that of the 1930 census The final adult standardization group included 1081 literate individuals (all white) ranging in age from 17 to 70 After this adult population had been selected, their educational status (educational level completed) was compared with that of the U S population at large, the assumption being that a similar distribution of educational levels of the standardization group would provide further evidence of its representative character In this respect, only moderate correspondence was found for on the whole the educational level of the standardization group is higher than that of the general population, although the range in both groups is from college graduates to illiterates

Since the Bellevue scale was to be standardized for some ages below

adult levels, it was necessary also to utilize a population of children For the selection of a representative population of the younger ages (sixteen and less), the chosen criterion was the age-grade distribution of pupils in the public schools of New York City About 1300 children were then tested in "representative and average" schools of New York City and about 200 in nearby communities in New York and New Jersey From these 1500, a number were selected (plus a small group of mental defectives from an institution) so as to yield an age-grade distribution that would fairly well approximate that of the New York City school population Thus, the final standardization population of the younger group included 670 white subjects, ranging in age from seven to sixteen years, and in grade placement from "ungraded" to twelve (plus continuation schools)

## VALIDITY

The nature of the content of the scale having been decided upon and norms of performance for a wide range of ages having been determined, the scale was subjected to several analyses intended to determine its validity

**Intercorrelations of subtests.** These coefficients are always necessary to provide data with regard to the presence or absence of a *g* factor For this scale, appreciable correlations would be required, as one aspect of its validity The author reports the following coefficients (each subtest correlated with every other subtest).—

Range of coefficients

15 (PE = 034), Object Assembly with Digit Span (N = 355; ages, 20–34)

.72 (PE = 026), Similarities with Comprehension (N = 355; ages, 20–34)

.37 to 72 is the range of the highest three-fourths of the intercorrelation coefficients (N = 590, ages, 20–49)

More significant are the coefficients found between scores on each subtest when correlated with total scores of *all other* parts of the scale[9] These coefficients were—

[9] For example, Information scores would be correlated with the total scores of all remaining parts If the scores on a given part—say, Information—were correlated with total scores of the test *including* the Information scores, it is obvious that the resulting coefficient would be in part *self-correlation* consequently it would be spuriously high

Range of coefficients

41 (PE = 029), for Object Assembly (N = 355, ages, 20–34)

73 (PE = 02), for both Similarities and Block Design (N = 590, ages, 20–49)

60 to 72 is the range of the highest three fourths of the coefficients (N = 590, ages, 20–49)

Additional evidence consistent with the presence of a general factor is provided by the following data showing the correlations between total scores on various combinations of subtests

Ages of subjects, 20 to 34 years
Number of subjects 355

83 (PE = 018), for total verbal subtest scores with total performance subtest scores

90 (PE = 014), for four subtests combining verbal and performance with a similar combination of four other subtests

The foregoing coefficients or intercorrelation cannot be said to be spuriously high, in spite of the fact that chronological age of the subjects varied considerably, for the subjects were adults, and in adult groups there is no continuous increase in score with successive ages as there is in the case of children who are still developing in mental capacity In calculating similar correlation coefficients for groups of children, it would be necessary to eliminate chronological age as a factor If this were not done the coefficient would appear to be higher than it actually should be, since both variables, being correlated would reflect the influence of a third factor namely, chronological age Similar statistical information obtained with a representative group of adolescents should have been provided in the process of standardization

**Correlation with schooling.** The reader will recall that beginning with Binet, the amount of schooling or quality of educational achievement —or both—were used as criteria of validity Hence, the ratings obtained on the Bellevue scale, for the adult standardization population were correlated with number of years of schooling, the coefficient being 64 (With mental defectives omitted, $r =$ 53 ) This coefficient of 64 falls within the range of coefficients commonly found when these two variables are correlated

**Correlation with teachers' judgments.** Teachers estimates of pupils' intelligence, rated on a six-point scale, were used as a validating criterion. For a group of adolescents (74 in number) in a trade school, the correlation coefficient between Bellevue IQ s and teachers' ratings was .52, for another group (45 in number) in a general high school, the coefficient was .43. The number of cases in each of these instances was too small to be very meaningful.

**Increase and decrease of scores with increasing age.** Tables of the scale show rises of mean scores until the age of 22.5 years, and thereafter slow decline. The rise in mean scores to age 22.5 is quite inconsistent with norms of other tests. The reasons for this have yet to be definitely determined. Was the standardization adult population more adequate than those of other tests? Or was it a selected group? Are the test materials more adequate and better scaled? Is the increase in norms an artifact due to methods of scoring?

**Constancy of means and standard deviation of IQ's at various ages.** The means are reasonably constant, ranging from 98.75 at the age interval 17–19 years, to 101.25 at age 10, with about sixty percent of the means falling between 100 and 101. The standard deviations range from 13.2 IQ points at age 10 to 16.85 at the age interval 50–59, while about half the standard deviations are between 14 and 15 points. (See page 136 regarding significance of constancy of means and standard deviations of IQ s.) These data indicate reasonably good satisfaction of this criterion.[10]

**Range and distribution of intelligence quotients.** The range of 1508 cases, ages 10 to 60, as shown in graph form is from about 45 to 145. The distribution curve for this group is slightly skewed in a negative direction,[11] that is, the IQ s fall within a somewhat narrower range above 100 than below. The author of the scale holds that the skewness does not bring into doubt the scale's validity, for he believes the widespread assumption of a symmetrical, bell shaped distribution (Gaussian) is a "mistaken belief." Many psychologists, however, would regard his curve of distribution as being a satisfactory approximation.

[10] The coefficients of variation $\left(\frac{SD}{M} \times 100\right)$ for the IQ s range from 13.04 at age 10 to 15.67 at the age interval 30–34.

[11] Wechsler interprets his curve as being "considerably skewed."

to the symmetrical curve which Terman assumed as necessary in standardizing the Stanford-Binet scales, and which Binet himself implied as desirable in the construction of his own scales

**Known groups.** In one study, two groups were differentiated on the basis of total scores namely, (1) a borderline group, having IQ s between 66 and 79, and (2) a mentally defective group, having IQ s between 50 and 65 The problem was to determine whether each of the eleven subtests contributes significantly to the differentiation of the two groups Since the mean scores on each of the subtests for the two groups did differentiate, and since the differences between the means were in the directions that should be expected, it was concluded that each subtest did contribute to the over-all differentiation, although Digit Span and Object Assembly contributed relatively little [12]

A second study used only the verbal subtests with naval recruits as subjects The problem here was to learn whether these subtests distinguish between (1) the mentally defective and the borderline, or (2) the borderline, the dull normal, and the normal The findings indicated that each of the verbal subtests contributed to the differentiations between these groups The Digit Span subtest in this instance, however, proved to be as effective as the others [13] This criterion of validity has still not been applied to the Bellevue scale with the same thoroughness as it has been applied to the Stanford Binet.

**Correlation with the Stanford-Binet.** In validating a new scale designed to test intelligence, it is common practice to correlate results obtained by means of the new device with ratings obtained on the Stanford Binet To do this is in effect to accept the Stanford Binet as a reasonably valid and reliable scale, and hence as one sound criterion of validity against which to evaluate the new scale

Using seventy-five cases (ages 14–16), the author of the Bellevue scale reports that Bellevue IQ s and Stanford-Binet IQ s when corre

[12] D Wechsler, et al "A Study of the Subtests of the Bellevue Intelligence Scale in Borderline and Mental Defective Cases" *American Journal of Mental Deficiency* Vol 45 pp 555–558 1941

[13] R. J Lewinski "Discriminative Value of the Subtests of the Bellevue Verbal Scale in the Examination of Naval Recruits" *Journal of General Psychology* Vol 31, pp 95–99 1944 Also H M MacPhee et al., "The Performance of Mentally Subnormal Rural Southern Negroes on the Verbal Scale of the Bellevue Intelligence Examination," *Journal of Social Psychology* Vol. 25, pp 217–229, 1947

lated yielded a coefficient of 82 (PE = 026) Six correlational studies of Bellevue and Stanford Binet ratings, made by others, have yielded coefficients ranging from 57 (PE = 04) to 93 (PE = 01). Although coefficients of partial correlation, with chronological age constant, are not given, it is doubtful if the age range in any of these studies is such as to have raised appreciably the size of the coefficient. In the two instances where the coefficients were below 80, the subjects were college freshmen, hence the relatively lower coefficients ( 57 and 62) may be attributable to either of the following conditions the fact that the group is relatively homogeneous, with resultant constriction of range and reduction in correlation, the fact that the tests are not as reliable at the upper extreme, so that errors of measurement make high correlations unlikely

Comparative studies, using the Stanford-Binet and the Bellevue, have not been as numerous as might have been expected in view of the wide use of both as clinical instruments Unfortunately, too—but understandably since the Bellevue's clinical application has been emphasized—a large percentage of the comparative studies have used hospital patients and clinic referrals as their subjects, to the neglect of school pupils and others who fall within categories of ' normal" behavior and adjustment

The available data, however, do show that there is very substantial correlation between these two scales, particularly when the Bellevue Full Scale IQ's and Verbal Scale IQ's are correlated with Stanford-Binet IQ's On the other hand, as would be expected, correlations between Stanford-Binet IQ's and Bellevue Performance Scale IQ's are only moderate Table 27 shows the distribution of correlation coefficients found in a number of representative studies [14]

Since correlation coefficients indicate *relative* agreement of paired scores, but not *absolute* differences between them, it is necessary, in comparing the Stanford-Binet and the Bellevue Scales, to know the extent of actual IQ differences existing between the correlated values On the whole it has been found that the differences are not very large In studies of retarded and mentally deficient persons (say, the lowest decile group), the Bellevue yields somewhat higher IQ's At the upper

[14] Correlational studies between the Bellevue and group scales have yielded varying results, the coefficients ranging from 39 ± 07 (PE), for Thorndike's CAVD test to 81 ± 04 (PE), for the Henmon Nelson Of six other coefficients, two are in the 70 s, two in the 60's, and two in the 50 s

level of mental ability, however (say, the highest decile group), the Stanford-Binet yields somewhat higher IQ's

Comparison of IQ s is complicated by the fact that age of testees is also a factor Taking the population samples *as a whole* (rather than only the extreme groups), the following are the general findings (1) Within the age range of approximately 10 to 19 years, the Stanford-Binet IQ's tend to be somewhat higher (2) From age 19 to about 35,

TABLE 27

*Coefficients of Correlation between the Stanford Binet and the Bellevue Scales (Frequencies)*

| r | *With Full Scale IQ* | *With Verbal IQ* | *With Performance IQ* |
|---|---|---|---|
| 90 | 5 | 2 | |
| 85 | 2 | 2 | |
| 80 | 1 | 1 | 1 |
| 75 | 2 | 1 | |
| 70 | | | 1 |
| 65 | | | 1 |
| 60 | 1 | 1 | |
| 55 | | | 1 |
| 50 | | | 2 |
| 35 | | | 1 |

the intelligence quotients tend to be about the same (3) Above the age of 35, the Bellevue intelligence quotients tend to be somewhat higher [15]

In the case of any given individual, therefore, comparison of Stanford Binet and Bellevue IQ s must take into account both factors ability level and chronological age In a given instance, a person's age and ability level may be such as to increase or decrease the difference between the IQ s obtained with the two instruments [16]

[15] Number (3) is just what one should expect in view of the method employed in calculating Bellevue IQ s The method is explained later in this chapter

[16] A definitive answer to the question of the comparability of the IQ s of the two scales will have to be based upon an investigation that is representative of the general population rather than heavily weighted with hospital and clinical subjects as is now the case Also such an investigation must approach the problem in two ways (1) by analyzing IQ s separately for each of the age groups and (2) by analyzing the IQ s separately at each of the ability levels for each of the age groups

On the whole, the correlation coefficients at hand, and other comparative data found between the two scales, are reasonably satisfactory and indicate that each has much in common with the other so far as concerns psychological functions being tested and ratings of ability levels.

**Prognostic efficiency.** The author of the Bellevue scale, as a final test of its validity, applies the pragmatic criterion. He states: "How do we know that our tests are 'good' measures of intelligence? The only honest reply we can make is that our experience has shown them to be so. If this seems to be a tenuous answer we need only remind the reader that it has been practical experience which has given (or denied) final validity to every other intelligence test. . . . Empirical judgments, here as elsewhere, play the role of ultimate arbiter. In any case, all evidence for the validity of a test, whether statistical or otherwise, is inevitably of an indirect sort and, in the end, cumulative rather than decisive." [17] In other words, it has been found by the author of the scale and by others that it works with reasonable satisfaction in clinical practice.

In evaluating the prognostic effectiveness of the Bellevue or the Stanford-Binet scale, or any other, one must bear in mind the age range for which a particular instrument is most effective. A blanket statement about a scale's diagnostic effectiveness is not warranted.

Furthermore, conclusions regarding the prognostic efficiency of a psychological scale are dependent upon the soundness of the clinical diagnoses with which the scale's findings are compared. Herein lies the major problem and weakness in attempts to determine prognostic efficiency of a psychological scale, for the determination of diagnostic clinical categories is difficult, often unreliable, and subject to the diagnostician's theoretical orientation. The major exception to this statement is the diagnosis of mental deficiency, for the determination of which a sound individual test of general intelligence is the most valid single instrument, when administered and interpreted by a qualified psychologist. For this purpose, both the Stanford Binet and the Bellevue scales have proved to be most valuable. It must be added at once, of course, that a diagnosis of mental deficiency is usually not made upon the basis of IQ and MA alone, although in some cases findings with a single test are so clear and unequivocal as to suffice.

[17] Wechsler, *op. cit.*, pp. 127–128.

## RELIABILITY

It is a noteworthy fact that very little research has been done on the reliability of the Bellevue scale. On the other hand, a great volume of material has been published on its clinical uses and interpretation, due very probably to the fact that the scale was developed and originally used primarily by a hospital staff of psychologists, and due also to the fact that it has been used most extensively by clinicians since its publication. The neglect of reliability studies, while research emphasis has been placed upon studies dealing with differential diagnosis, intellectual deterioration, intellectual changes under treatment, etc., is regrettable, for the basic soundness of an instrument should be studied at least collaterally with its application.[18]

Wechsler's manual itself reports only very meager and inadequate data on reliability: namely, 52 individuals retested at intervals of one month to one year, with the results shown in Table 28.

TABLE 28

*Retest Correlation Coefficients for the Bellevue Scale*[19]

| *Ages* | *N* | *Rho* * | *P E* |
|---|---|---|---|
| 10–13 | 32 | .94 | .013 |
| 20–34 | 20 | .94 | .018 |

*Rank order correlation coefficient

Since publication of the foregoing data, even the few available reliability studies have dealt almost exclusively with abnormal subjects, principally psychoneurotics and schizophrenics.[20] Table 29 shows the range of correlation coefficients found with such groups for each of the subtests and IQ scales.

[18] The ready use of a new and promising clinical instrument is understandable and justifiable since clinicians confronted by immediate pressing and persistent problems of living human beings cannot wait until experimental research has subjected the instrument to thoroughgoing tests of reliability and validity. In using an instrument, however, it is essential that we give due consideration to its limitations and to unanswered or only partially answered questions about it.

[19] From Wechsler, *op. cit.*, p. 133. (By permission.)

[20] Because of the instability of such persons, they are not the most suitable subjects to use for the study of the inherent stability of a measuring instrument.

Examination of Table 29 shows (1) that there is considerable variation in reliability among the subtests and that, in general, their reliability is appreciably below that of the scale as a whole, (2) that in all but one of the reports (in which the subjects were schizophrenics and $r = 55$) full scale reliability appears to be reasonably satisfactory, considering the instability of the groups used (Other $r$'s were· .87, 84, 84, 87, 89, 90 )

TABLE 29

*Test Retest Reliability of the Bellevue Scale Reported for Abnormal Groups (7 Studies)*

| *Subtest* | *Range of Coefficients* |
|---|---|
| Information | 56–99 |
| Comprehension | 12–78 |
| Digit Span | 59–77 |
| Arithmetic | 68–87 |
| Similarities | 38–93 |
| Vocabulary | 90–93 |
| Picture Arrangement | 49–86 |
| Picture Completion | 32–89 |
| Block Design | 65–87 |
| Object Assembly | 31–79 |
| Digit Symbol | 34–91 |
| Verbal IQ | 76–91 |
| Nonverbal IQ | 52–94 |
| Full Scale IQ | 55–90 |

From published reports, it appears that only one investigation, using a fairly adequate sampling of individuals, has been devoted to reliability of the Bellevue when administered to "normal" subjects [21] The test-retest method was employed The age range was 20 to approximately 50 years One group of 60 subjects was retested after a one-week interval, another group of 60 persons after a four-week interval, a third group of 38 subjects after a six month period The major findings were the following

[21] G F Derner M Aborn, and A H Canter, "The Reliability of the Wechsler Bellevue Subtests and Scales" *Journal of Consulting Psychology*, Vol 14 1950, pp 172–179 See also W B Webb and H DeHaan, 'Wechsler Bellevue Split Half Reliabilities in Normals and Schizophrenics," *ibid*, Vol 15, pp 68–71, 1951 The split half method is an inappropriate method to use with most of the Bellevue scale

The mean score for every subtest and for the total scale increased for all three groups

Increases in scores tend to be somewhat smaller as the retest interval is increased

The smallest average increase was 0 3 points in weighted score for comprehension retest (after four weeks)

The largest average increase was 2 8 points in weighted score for picture arrangement retest (after one week)

Largest average increases (2 or more points in weighted score) were found for picture arrangement and object assembly

Smallest average increases (less than one point in weighted score) were found for information, comprehension, and similarities

Average changes in IQ's were verbal scale, 4 4 points, nonverbal IQ, 9 1 points, full scale IQ 7 6 points

Retest correlations and standard errors of measurement[22] for all subtests and the three IQ s are shown in Table 30 It will be noted

TABLE 30

*Test Retest Correlations and Standard Errors of Measurement for the Bellevue Scale*[23]
($N = 158$)

| *Subtests* | *Correlations* | *S E meas* |
|---|---|---|
| *Information* | 86 | 68 |
| Comprehension | 74 | 1 21 |
| Digit Span | 67 | 1 68 |
| Arithmetic | 62 | 2 06 |
| Similarities | 71 | 1 22 |
| Vocabulary | 88 | 73 |
| Picture Arrangement | 64 | 1 82 |
| Picture Completion | 83 | 95 |
| Block Design | 84 | 1 10 |
| *Object Assembly* | 69 | 1 31 |
| Digit Symbol | 80 | 1 06 |
| Verbal IQ | 84 | 3 96 |
| Nonverbal IQ | 86 | 4 49 |
| Full Scale IQ | 90 | 3 29 |

[22] For interpretation of standard error of measurement see Chapter 1, p 17
[23] From G F Derner, et al, *op cit*

that, for the subtests, four of the coefficients are in the 60's (very low reliability), two are in the 70's (low reliability), five are in the 80's (satisfactory reliability for a subtest)

On the whole, it appears from this study, using subjects within the "normal' range of personality and behavior, that the reliabilities of the Bellevue scale are not high enough in six of the eleven subtests to warrant their use independently This is a significant finding, especially as it relates to the use of the subtest profile for diagnostic purposes (discussed in Chapter 15) The reliability coefficients for the Verbal and Nonverbal IQ ratings are reasonably satisfactory, though not as high as some psychologists would demand The full scale IQ reliability, however, is quite satisfactory The standard errors of measurement, it will be noted, conform rather closely with the correlations The general conclusion, then, is that while the subtests individually do not show a high degree of reliability, *the scale as a whole yields reasonably reliable results,* as indicated by the correlation coefficients and the standard errors of measurement

These results indicate, again, the significance and value of measuring a number of functions, for not only does such a measure yield a more representative index, but there is a greater probability that daily or periodic fluctuations in performance will be compensated for through the testing of a number of representative functions

## SCORING AND IQ CALCULATION

**Scoring.** All parts of this scale are scored on a point basis For some subtests, the earned raw score is simply the number correct, each item being scored either plus or minus (e g , information) Or, as in the case of comprehension and similarities, the score for each item is 0, 1, or 2, depending upon quality of the response In other parts, as in arithmetical reasoning and block design, the earned raw score is based not only upon correct responses but upon the time taken to solve the problem Thus, the factor of speed of performance is involved in sections of this scale, especially in nonverbal subtests

The *raw score* for each subtest is first obtained by simple addition of the credits on the items in that part This raw score is then converted into a weighted score (a type of standard score), by means of a conversion table The purpose of this conversion is the customary one of placing all subtest scores on a comparable basis The weighted

scores for all parts of the scale are then added to obtain the full score upon which the "full scale IQ" is based. Also, the weighted scores of only the six verbal parts are added to get the verbal score, upon which the "verbal scale IQ" is based. Similarly the weighted scores of the five performance tests are added to get the performance score and "performance scale IQ."

The following well known formula was used for equating each subtest's raw score into *weighted scores*

$$X_2 = M_2 + \frac{SD_2}{SD_1}(X_1 - M_1),$$

in which

$M_2$ = an arbitrarily assigned mean (10)
$SD_2$ = an arbitrarily assigned standard deviation (3)
$X_2$ = the weighted score to be found
$M_1$ = the mean of the subtest's raw score
$X_1$ = *the particular raw score to be converted to a weighted score*

What this formula does is (1) assign an arbitrary and *uniform mean score* to all subtests, (2) multiply each individual score's deviation from its mean by a *constant ratio* (3) add the result to or subtract it from the assigned mean.[24] By using this formula, scores are so converted that each individual maintains his relative status on each subtest. And in the case of any given person's subtest scores, differences between scores will be attributable, theoretically, to differences in his performance level rather than to differences in the weighting of each subtest in the total. It is thus possible to vary the number of items in each of the several subtests without giving any of them too little or too much weight in the total score upon which the IQ is based.

On the Bellevue scale, the reason for converting raw scores into weighted scores is this: the possible maximum raw scores vary in the several subtests of the scale. If, therefore, the raw scores were simply

[24] The best way for the student to see how this formula works is to substitute several sets of values in it and to observe the outcome. The logic of the process will then be more readily apparent. For example, if the two following sets of values are substituted in the formula, the process will be clear. Assume the following data for one subtest: mean score = 12, SD = 4, $X$ = 15, while for a second subtest *the corresponding values are 24, 8, and 30. The results will* show the same weighted score for both subtests (12.25) because the individual's relative status on each subtest was identical with that on the other subtest, even though the raw scores differ.

added to obtain an individual's rating on the scale, each of the parts would carry a different weight in the total, each part would have the possibility of contributing differently to the final result—some more heavily than others The raw-score units of one part of the scale would not have the same significance as those of other parts If this were the case, then implicit in the scoring would be the assumption that certain of the psychological functions being tested should be regarded as more important than others in the total score and in getting an index of intelligence The Bellevue scale, however, is scored on the principle that all the functions tested are equally important, hence, the part-scores should be equally weighted so that each part may contribute as much to the total as any other part

**Calculating the IQ.** After the verbal score, performance score, and full score have been obtained (in converted units), the three corresponding intelligence quotients are found in tables prepared for that purpose In calculating IQ s for the Bellevue scale, the *basic principle* used differs from that in the Stanford-Binet and in most other scales The *principle* employed is that an individual s intelligence quotient should be determined by the relative extent to which his weighted score (full, verbal, or nonverbal) deviates from the mean weighted score of *his own age group* The detailed method of actually determining Bellevue intelligence quotients involves several steps and assumptions

First, the mean weighted score and the standard deviation for each age level are calculated Second, the weighted scores at each age level are converted into standard scores (*Z* scores See Chapter 2, page 45) Then, third, it is assumed that a value of 6745 of a standard score shall be equated with an IQ of 90[25] Hence, this assumption means that the probable error ( 6745 SD) of Bellevue IQ s is arbitrarily set at 10 points, and the standard deviation of IQ's at approximately 15 points, since the PE equals 6745 of the SD[26] In other words, according to this device, in a "normal," symmetrical distribu-

[25] The reason why 6745 *Z* is taken is this the index called "probable error" (PE) is 6745 of a standard deviation The standard score, it will be recalled is an index given in terms of the standard deviation The PE therefore is convenient and easily calculated when the standard scores are known

[26] This device makes Bellevue IQ s approximate the most probable SD of the Stanford Binet distribution of intelligence quotients namely, about 16 points

scores for all parts of the scale are then added to obtain the full score upon which the "full scale IQ" is based Also, the weighted scores of only the six verbal parts are added to get the verbal score, upon which the "verbal scale IQ' is based Similarly the weighted scores of the five performance tests are added to get the performance score and "performance scale IQ"

The following well known formula was used for equating each subtest's raw score into *weighted scores*

$$X_2 = M_2 + \frac{SD_2}{SD_1}(X_1 - M_1),$$

in which

$M_2$ = an arbitrarily assigned mean (10)
$SD_2$ = an arbitrarily assigned standard deviation (3)
$X_2$ = the weighted score to be found
$M_1$ = the mean of the subtest's raw score
$X_1$ = the particular raw score to be converted to a weighted score

What this formula does is (1) assign an arbitrary and *uniform mean score* to all subtests, (2) multiply each individual score's deviation from its mean by a *constant ratio* (3) add the result to or subtract it from the assigned mean[24] By using this formula, scores are so converted that each individual maintains his relative status on each subtest And in the case of any given person's subtest scores, differences between scores will be attributable, theoretically, to differences in his performance level rather than to differences in the weighting of each subtest in the total It is thus possible to vary the number of items in each of the several subtests without giving any of them too little or too much weight in the total score upon which the IQ is based

On the Bellevue scale, the reason for converting raw scores into weighted scores is this the possible maximum raw scores vary in the several subtests of the scale If, therefore, the raw scores were simply

[24] The best way for the student to see how this formula works is to substitute several sets of values in it and to observe the outcome The logic of the process will then be more readily apparent For example if the two following sets of values are substituted in the formula the process will be clear Assume the following data for one subtest mean score = 12, SD = 4, $X$ = 15 while for a second subtest the corresponding values are 24, 8, and 30 The results will show the same weighted score for both subtests (12 25) because the individual s relative status on each subtest was identical with that on the other subtest even though the raw scores differ

added to obtain an individual s rating on the scale each of the parts would carry a different weight in the total each part would have the possibility of contributing differently to the final result—some more heavily than others The raw score units of one part of the scale would not have the same significance as those of other parts If this were the case then implicit in the scoring would be the assumption that certain of the psychological functions being tested should be regarded as more important than others in the total score and in getting an index of intelligence The Bellevue scale however, is scored on the principle that all the functions tested are equally important hence the part scores should be equally weighted so that each part may contribute as much to the total as any other part

**Calculating the IQ** After the verbal score performance score and full score have been obtained (in converted units), the three corresponding intelligence quotients are found in tables prepared for that purpose In calculating IQ s for the Bellevue scale the *basic principle* used differs from that in the Stanford Binet and in most other scales The *principle* employed is that an individual s intelligence quotient should be determined by the relative extent to which his weighted score (full verbal or nonverbal) deviates from the mean weighted score of *his own age group* The detailed method of actually determining Bellevue intelligence quotients involves several steps and assumptions

First the mean weighted score and the standard deviation for each age level are calculated Second the weighted scores at each age level are converted into standard scores (*Z* scores See Chapter 2 page 45) Then third it is assumed that a value of 6745 of a standard score shall be equated with an IQ of 90[25] Hence this assumption means that the probable error ( 6745 SD) of Bellevue IQ s is arbitrarily set at 10 points and the standard deviation of IQ s at approximately 15 points since the PE equals 6745 of the SD[26] In other words according to this device in a normal symmetrical distribu

[25] The reason why 6745 *Z* is taken is th s the index called probable error (PE) is 6745 of a standard deviation The standard score it will be recalled is an index given in terms of the standard deviation The PE therefore is conven ent and easily calculated when the standard scores are known

[26] This device makes Bellevue IQs approx mate the most probable SD of the Stanford Binet d stribut on of intell gence quotients namely about 16 points

tion of Bellevue IQ s, fifty percent of the IQ values will fall between 90 and 110 (±1 PE), since the PE sets the limits of the middle fifty percent of the scores of a distribution, and about two thirds (68 26 percent) of the values will fall between 85 and 115 (±1 SD)

On the basis of the foregoing assumption, it is possible to assign an IQ value to any weighted standard score In effect, what the procedure amounts to is that the standard score is converted into a probable error value, this value is multiplied by 10, and the result is added to or subtracted from 100, depending upon whether the standard score is plus or minus The calculations are shortened and facilitated by a formula for the purpose[27] The intelligence quotients found in this manner should be called "deviation IQ s"

**Criticism of the Bellevue Method.** When this method of determining intelligence quotients is used, the following points should be noted This method consistently compares (by means of IQ s) an individual s test score with the mean for his own chronological age group up to 60 years, whereas on the Stanford Binet and other scales a person whose CA is greater than the age at which "average adult MA" is reached is compared with a given and fixed maximum age level Thus it will be recalled that on the 1916 Stanford Binet, in calculating the IQ of a person 16 years of age or older, the maximum value in the denominator is 16 (MA/16) regardless of his actual age When using the 1937 Stanford Binet, the maximum denominator is 15 In other words, the method used with the 1937 Stanford-Binet relates the test performance of a person above age 15 to the performance level of an average group at a specified maximum age (15), the Bellevue method, at all ages up to sixty, relates a person's test performance to the average performance of *his own* age group Also, then, the Bellevue method, for the purpose of IQ calculation, obviates the necessity of determining age of "average adult MA"—always a difficult problem and as yet uncertain

One weakness of the Bellevue method of calculating IQ is this the same, or constant, objective performance on the test (i e , the score) will rate an individual higher with increased age after the maximum, since his constant score will be compared with steadily declining age norms For example, a person who earns an IQ of 60 at age 20, main-

[27] See Wechsler *op cit* pp 219 220

taining a constant score to age 50, would have an IQ of 78 at age 50 To maintain an IQ of 60 there would need to be an average rate of decline consistent with an IQ of 60

Another point concerns the question of when an individual's mental level, *as measured by the tests,* begins to decline This, too, presents a question as yet not definitively answered There is also the question of the *rate* of decline in functions being tested during each of the several stages of adulthood The Bellevue IQ's would not provide any evidence toward answers to either of these questions But regardless of specific answers to these two questions, data show that with the Bellevue scale there is some slight decline in average *test scores* between the ages of approximately 25 and 45, and that rate of decline in average test scores increases thereafter[28] Obviously, therefore, after the period of decline (however moderate) sets in, as shown by the tests, an individual's IQ will begin gradually to decline if the Stanford-Binet method is used, whereas if the Bellevue method of calculating IQ is used, an individual's rating will decline only if he loses ground *with reference to the average of his own age group,* rather than with reference to a more or less hypothetical "average adult level" If his losses are at the same general rate as those of his age group, his Bellevue IQ will remain relatively constant If, however, his losses are less than the general rate, his Bellevue IQ will rise

## SPECIAL FEATURES OF THE BELLEVUE SCALE

The Bellevue scale provides a scheme for calculating a "deterioration quotient" based on the premise that certain types of tested mental processes decline more rapidly than do other types, and that the difference between rates of decline, as between these two types, in the case of any given person indicates his relative degree of deterioration In other words, there are certain tested functions that hold up with age and others that do not hold up with age This index

---

[28] The reader should note that we have emphasized decline in test score" This does not necessarily mean that on the whole a person becomes progressively "less intelligent" even before the effects of senescence become apparent While it is true that there is some loss in average test scores after about age 25, it is also true that some mental traits, as yet unmeasured by intelligence tests, increase in effectiveness through an extended period of adulthood and more than compensate for losses in the processes measured by current scales This view is borne out by the facts regarding ages of maximum achievement of scholars, scientists writers, and artists

will be dealt with in more detail in a later chapter together with other tests devised to measure deterioration of mental abilities

A second feature of the Bellevue scale is its emphasis upon scatter analysis —that is analysis of an individual s performance on the several parts of the scale for the purpose of facilitating clinical analysis of the subject s performance Such analysis may lead to diagnostic inferences concerning personality characteristics and behavior disorders due to organic brain disease psychosis psychoneurosis adolescent psychopathy and mental deficiency Here again this application of the scale and clinical evidence supporting and contrary will be presented in a subsequent chapter on clinical uses and interpretations of tests

## CRITICISMS AND EVALUATIONS

The Bellevue scale is being quite widely used in the measurement of adult intelligence especially in psychological clinics Implicit in the wide use of a testing instrument is endorsement and acceptance of the scale at least as one of the more satisfactory of those available at the time

**Was the population sample adequate?** Since the 1751 persons upon whom this scale was standardized were from New York City and nearby communities the adequacy of the population sample may be and has been seriously questioned It would be highly desirable for the author of this scale and his associates to assemble and publish a frequency distribution of scores and intelligence quotients obtained with normal groups in various sections of the country These data should be separately presented for each of the several age levels they should be analyzed for socio-economic differences and sex differences to determine if these are significant

The occupational distribution of the standardization population was a moderate approximation to the 1930 census which was taken as the basis for the selection of the population sample Furthermore the validity of occupational regrouping in the process of standardization may be questioned [23] Also in the standardization population there were more persons from the upper educational levels than there were in the general population

[23] See Wechsler *op cit* p 111

In view of the foregoing considerations, the norms of the subtests and the totals should not be regarded as final

**Are the subtests a variety of disconnected types of tests?** With the exceptions of object assembly and, to a lesser extent, picture arrangement, the intercorrelations of the subtests are all significant and marked This would suggest that the Bellevue subtests are measuring one or more common factors to an appreciable extent The *inter*correlations between verbal and nonverbal subtests are not, on the whole, as high as *intra*correlations within each of these categories, but with the exceptions noted the *inter*correlations are indicative of a common factor or common factors

This scale has not been adequately analyzed factorially One analysis, however, finds that a "first factor" (general factor) accounts for from 27 percent to 50 percent of individual differences in scores, the weight of the general factor varying apparently at different age levels [30] The high correlations between the Stanford-Binet and the Bellevue scales indicate that the two instruments have much in common as regards the psychological functions measured by them And, it will be recalled, the Stanford-Binet has been found to measure primarily a general factor The question of factors as determined from analysis of Bellevue scale results themselves, is, however, one that should be subjected to further comprehensive investigation [31]

**Are the verbal subtests culturally unfair to some persons?** The answer to this question is the same as the one given for the Stanford-Binet In addition, it may be said that in this scale, the verbal materials in Comprehension, Similarities, and Arithmetic are stated in such terms as place very little premium upon educational or other

[30] I Lorge *Third Mental Measurement Yearbook*, O K Buros, editor, New Brunswick N J Rutgers University Press, 1949, p 393 See also B Balinsky, *An Analysis of the Mental Factors of Various Age Groups from Nine to Sixty*, Genetic Psychology Monographs, Vol 23, pp 191-234, 1941

[31] The need of such research is still present even though a few factorial studies have been published recently These studies are, unfortunately, limited to atypical groups See J Cohen "A Factor Analytically Based Rationale for the Wechsler Bellevue," *Journal of Consulting Psychology* Vol 16 pp 272-277, 1952 This report deals with psychoneurotics, schizophrenics and brain-damaged individuals The findings, therefore are not as universally applicable as the title implies Also J E Birren, "A Factorial Analysis of the Wechsler-Bellevue Scale Given to an Elderly Population" *ibid*, Vol 16, pp 399-405, 1952 Since the subjects of this report were between 60 and 74 years of age, the findings can hardly be regarded as representative

cultural advantages. Like all tests of information and vocabulary, performance on these in the Bellevue is dependent, in part, upon opportunity to learn, whether in school, home, or through the intellectual exploitation of all aspects of one's environment.

**Are some of the test items obsolete?** Items in any and all tests must be reviewed periodically for possible obsolescence. In the case of the Bellevue, a number of the items, especially in comprehension, information, vocabulary, picture arrangement, and picture completion should be re-examined and re-evaluated. Also for a number of items the satisfactory, partially satisfactory, and unsatisfactory responses should be reviewed and revised in the light of responses that have been obtained since the scale's original publication.

**Is the factor of speed important in the Bellevue scale?** Unlike the Stanford-Binet, in which very few test items are timed, Bellevue scores are significantly affected by the speed factor. Speed of performance yields additional credits in the following subtests: arithmetic, picture arrangement, object assembly, digit symbol, and block design. Thus, in the total score, speed of work is combined with power (or ability level). Although in general speed and power are highly correlated, it is also a fact that response time slows down with age. Thus, since the Bellevue scale is designed for adults, to an important degree it measures, especially in later adult years, decline in speed of response and not necessarily decline in power. This factor must be kept in mind when, in a later chapter, we consider the "decline" of abilities and the suggested "deterioration" index.

**Do the nonverbal subtests involve visual acuity?** Although no experimental data are available in answer to this question, not a few users of the Bellevue scale have observed that visual acuity might be a factor in some instances. The subtests most likely to make some demands upon visual acuity are picture arrangement and picture completion. And, of course, color blindness must be eliminated as a factor in the block design subtest.

**Are the reliability coefficients satisfactory?** Available data indicate that the total verbal scores and the total performance scores have a fairly satisfactory degree of reliability; while the full scale scores have a degree high enough to satisfy the standards of most psychologists. The reliabilities of the subtests, however, are not high enough, when

taken individually, to be used uncritically for differential diagnosis or for evaluating deterioration of mental ability, since accuracy in both of these matters depends upon accuracy of retest results More research is necessary on the several reliabilities for a normal population rather than for hospital and clinical cases

**Should the Bellevue scale use the IQ?** There is warranted criticism of the use of the term "intelligence quotient for the index derived with this scale, because the formula employed really changes the conception of the IQ and thereby confuses the meaning of the term Meanings of scientific terms are established by priority and usage, and usage had established a conception of the IQ as developed by the Binet revisions and by group tests which followed the same basic principle as regards the maximum denominator in the formula (IQ = MA/CA) Since the IQ has been and is being used with the Bellevue scale, it is necessary to bear in mind that it is a ' deviation IQ

**Should the verbal and the performance scores be combined?** The tables of norms for this test show that the maximum verbal score norm is reached at the age of 22 5 years while the maximum performance score norm is attained at the 16 5–18 years Maximum full scale norm is found at the age of 22 5 years In view of these differences we may question the wisdom of combining verbal and performance scores, particularly after the age of 18 when one group of functions (performance) no longer develops differentially, while the other group of functions (verbal) does so develop It is quite possible that these two sets of tests and functions are sufficiently different after the age of 18 so that they should be separately scored It is also possible that the value of the scale is reduced by combining the two sets of subtests Furthermore, it is possible that some of the inadequacies of the Bellevue Scale findings may be attributable to combining the two types of subtests

**Is the Bellevue scale clinically useful?** Judging from its widespread use in clinics and hospitals and from the empirical judgments of many clinicians, it appears that this scale has been of considerable value Used with scientific judgment and with knowledge of its limitations and its tentativeness in some aspects, the Bellevue scale can be a very useful instrument for estimating intelligence of adolescents and adults An important qualification must be added namely, that this

cultural advantages Like all tests of information and vocabulary, performance on these in the Bellevue is dependent, in part, upon opportunity to learn, whether in school, home, or through the intellectual exploitation of all aspects of one's environment

**Are some of the test items obsolete?** Items in any and all tests must be reviewed periodically for possible obsolescence In the case of the Bellevue, a number of the items, especially in comprehension, information, vocabulary, picture arrangement, and picture completion should be re-examined and re-evaluated Also for a number of items the satisfactory, partially satisfactory, and unsatisfactory responses should be reviewed and revised in the light of responses that have been obtained since the scale's original publication

**Is the factor of speed important in the Bellevue scale?** Unlike the Stanford-Binet, in which very few test items are timed, Bellevue scores are significantly affected by the speed factor Speed of performance yields additional credits in the following subtests arithmetic, picture arrangement, object assembly, digit symbol, and block design Thus, in the total score, speed of work is combined with power (or ability level) Although in general speed and power are highly correlated, it is also a fact that response time slows down with age Thus, since the Bellevue scale is designed for adults, to an important degree it measures, especially in later adult years, decline in speed of response and not necessarily decline in power This factor must be kept in mind when, in a later chapter, we consider the "decline" of abilities and the suggested "deterioration" index

**Do the nonverbal subtests involve visual acuity?** Although no experimental data are available in answer to this question, not a few users of the Bellevue scale have observed that visual acuity might be a factor in some instances The subtests most likely to make some demands upon visual acuity are picture arrangement and picture completion And, of course, color blindness must be eliminated as a factor in the block design subtest

**Are the reliability coefficients satisfactory?** Available data indicate that the total verbal scores and the total performance scores have a fairly satisfactory degree of reliability, while the full scale scores have a degree high enough to satisfy the standards of most psychologists The reliabilities of the subtests, however, are not high enough, when

taken individually, to be used uncritically for differential diagnosis or for evaluating deterioration of mental ability, since accuracy in both of these matters depends upon accuracy of retest results More research is necessary on the several reliabilities for a normal population rather than for hospital and clinical cases

**Should the Bellevue scale use the IQ?** There is warranted criticism of the use of the term "intelligence quotient" for the index derived with this scale, because the formula employed really changes the conception of the IQ and thereby confuses the meaning of the term Meanings of scientific terms are established by priority and usage, and usage had established a conception of the IQ as developed by the Binet revisions and by group tests which followed the same basic principle as regards the maximum denominator in the formula (IQ = MA/CA) Since the IQ has been and is being used with the Bellevue scale, it is necessary to bear in mind that it is a 'deviation IQ'

**Should the verbal and the performance scores be combined?** The tables of norms for this test show that the maximum verbal score norm is reached at the age of 22 5 years, while the maximum performance score norm is attained at the 16 5–18 years Maximum full scale norm is found at the age of 22 5 years In view of these differences we may question the wisdom of combining verbal and performance scores, particularly after the age of 18, when one group of functions (performance) no longer develops differentially, while the other group of functions (verbal) does so develop It is quite possible that these two sets of tests and functions are sufficiently different after the age of 18 so that they should be separately scored It is also possible that the value of the scale is reduced by combining the two sets of subtests Furthermore, it is possible that some of the inadequacies of the Bellevue Scale findings may be attributable to combining the two types of subtests

**Is the Bellevue scale clinically useful?** Judging from its widespread use in clinics and hospitals and from the empirical judgments of many clinicians, it appears that this scale has been of considerable value Used with scientific judgment and with knowledge of its limitations and its tentativeness in some aspects, the Bellevue scale can be a very useful instrument for estimating intelligence of adolescents and adults An important qualification must be added namely, that this

scale is not adequate to measure and differentiate the highest levels of ability, at least the upper five percent of the population

The Bellevue scale is a valuable addition to other testing and diagnostic devices, such as the Stanford Binet, the Arthur Performance Scale, the Babcock test, and others which will be presented Between most of these scales there are significant correlations, the exclusive clinical effectiveness of one or the other has yet to be established For the present, and no doubt in the future, clinicians (in schools hospitals, and elsewhere) will use a given scale or a combination of scales as occasion demands and as their clinical insights suggest

Some psychologists give considerable weight to the fact that the Bellevue is so constructed that it is possible to analyze an individual s scores in terms of his variations (consistency or inconsistency) on the several parts of the scale, especially since attempts have been made by the author of the scale and by others to specify the psychological functions beiog tested by each of the several parts

The validity of the Bellevue in identifying persooality and behavior disorders has yet to be unequivocally demonstrated In spite of the fact that this is the area in which most of the evaluative studies of the scale have been made, clinical findings are by no means definitive

In attempting to diagnose personality and behavior disorders on the basis of the pattern or profile of scores on the Bellevue or other scales it must be remembered also that different educational backgrounds and cultural factors, quite unrelated to personality and behavior disorders, could account, to some degree, for an individual s inconsistency of performance on the several parts of the scale It has been found too, that individual variations in interests, as distinguished from personality disorders, find expression in different patterns of mental activities and are reflected in subtest variations However, the results obtained by means of the Bellevue, plus clinical experience and acumen provide a valuable combination for study of individual differences and individual mental functioning

**General comment.** Since its appearance in 1939, the Bellevue scale has been widely reviewed and evaluated [32] Judgments have varied

[32] See O K. Buros ed *op cit* pp 386–398 and Buros *The Fourth Mental Measurement Yearbook* Highland Park N J The Gryphon Press 1953 pp 473–476

from enthusiastically uncritical acceptance to destructively critical rejection Some critics have concentrated, and justifiably, upon unwarranted assumptions and statistical inadequacies, to the exclusion of other and positive aspects At the other extreme are the critics who have ignored the scale's defects and limitations while concentrating upon and lauding its practical value Most evaluations, however, have been moderate in that they have pointed out the contributions, values, and possibilities of the scale, while clearly indicating its defects and doubtful assumptions Experimental research and competent practical application can proceed simultaneously, each facilitating the other [33]

## THE 1955 REVISION OF THE BELLEVUE SCALE

Early in 1955 a revised edition of this scale is scheduled for publication This new edition, it appears, will meet some of the adverse criticisms directed against the original scale The revised version (to be known as the Wechsler Adult Intelligence Scale) does not introduce any new principles in its content, construction, organization, scoring, or IQ derivation The main changes are in the revision of some content, extension of the population sample, and in improvement in directions for administering and scoring Some of the major revisions (supplied by The Psychological Corporation prior to publication of the scale) are the following

**Content.** Range of difficulty has been extended, chiefly downward in order to assure a score for the lower level of mentally deficient subjects Upward extension in difficulty has been slight Progression of difficulty from item to item has been improved Obsolete items have been replaced Items having poor "item validity" and those overlapping others in content have been replaced, as have those that were ambiguous Illustrations in the picture completion subtest have been

[33] Some of the other questions asked in connection with the Stanford Binet might appropriately be asked about the Bellevue But since the principles involved and the replies would be the same they have not been repeated

Form II of the Bellevue scale has been made available (New York The Psychological Corporation 1946) This form will not be discussed because (1) it is presumably identical with Form I in respect to underlying principles and types of test materials (2) the standardization information provided is too limited and the data provided indicate that Form II does not meet the criteria necessary if it is to be used as an alternate scale that is equal or very nearly equal means, deviations, and distributions obtained with the same population sample

more clearly drawn The vocabulary subtest has been revised so as to produce a fairly normal distribution of scores for a representative sample of the population Maximum scores in verbal and nonverbal subtests and in the full scale are reached by the 25–34 year age group

**Population sample** Norms are based upon a sample of 1700 per sons 850 of each sex selected from four major geographic areas The subjects ranged in age from 16 to 64 years The age range was di vided into seven age groups within each of which the numbers were proportioned according to the 1950 U S census with respect to geographic area race (white and nonwhite) occupation urban rural and years of formal education Supplementary data were also obtained for a sample of older persons (N = 352) above 65 years of age

**Reliability and Validity** For the separate subtests the reliability estimates range from 66 (picture arrangement) to 96 (vocabulary) Reliabilities for the three IQ s are verbal scale 96 performance scale 93 full scale 97 Intercorrelations among the eleven subtests range from 30 to 85 while the correlation coefficient for total verbal scores *vs* total performance scores (ages 18 24) is 77

## THE WECHSLER INTELLIGENCE SCALE FOR CHILDREN (1949)[34]

**Description** This scale for children from five through fifteen years of age is built on the same principles and in the same form as the Bellevue scale for adolescents and adults verbal subtests per formance subtests a verbal IQ a performance IQ and a full scale IQ

The subtest types are identical with those of the older scale with the following exceptions digit span is made optional an opt onal maze test has been added and in place of d git symbol, a coding test has been substituted in which various lines in varied positions (single double circle) are associated with geometric figures (star circle triangle cross rectangle)

**Standardization Population** The scale was standardized on a sample of 100 boys and 100 girls at each of the eleven age levels each child being tested within one and one half months of his mid year In ef

[34] New York, The Psychological Corporat on.

fect this means that children were selected at the half way mark between birthdays, and half way was defined as being between four and a half months and seven and a half months (excepting the feebleminded, nearly all of whom were within two months of their midyear)

Selection of the 2200 children was based upon (1) rural urban residence, (2) father s occupation, and (3) geographic area The proportions in these sampling factors were based upon U S Census data for 1940, with some adjustment for the shift of population toward the West ' In the final selection of the standardization sample, geographic area percentages are reasonably well satisfied, urban rural percentages less well and father s occupation percentages moderately[35] On the whole the standardization group satisfies the principles of sampling better than did the population sample used for the Bellevue adult scale Still 2200 children distributed among eleven age groups and over four very wide geographic areas (New England and Middle Atlantic States, North Central States, South Atlantic and South Central States Mountain and Pacific States) are but a small handful Extensive experimental use of this scale will be necessary to determine the adequacy of the norms based upon the population sample used in standardization

**Reliability Data**[36] Reliability coefficients were found for three age groups (7½, 10½, 13½) the number being 200 in each The findings are summarized in Tables 31 and 32 It will be noted, from these data, that the subtest reliability coefficients vary markedly and are, on the whole, only moderate in magnitude The IQ reliabilities, however, being from 86 to 96, fall within the range that is generally acceptable These data demonstrate again the necessity of distinguishing between reliability of part of a scale and that of the whole scale

[35] For detailed standardization see Wechsler *op cit* and H Seashore et al 'The Standardization of the Wechsler Intelligence Scale for Children *Journal of Clinical Psychology* Vol 14 pp 99 110 1950

[36] The split half technique was used to calculate reliability except in the case of coding and digit span For the former results of coding tests A and B were used since this is essentially a speed test For digit span scores on digits forward were correlated with scores on digits backward—a very questionable procedure since the two do not involve identical processes Also some question may be raised regarding the appropriateness of the split half method for some of the other subtests The test retest method is much to be preferred for a scale of this type

The standard error of measurement indicates the range of score within which the chances are approximately two to one that a subject's "true" score will fall in that particular subtest. Thus, the standard error of 1.20 for 7½-year-olds on picture arrangement indicates that the probabilities are two to one that an individual's "true" score on this subtest is within 1.20 points of his obtained weighted score. Likewise the standard error of 4.25 IQ points (full scale) for 7½-year-olds indicates that the probabilities are about two to one that an

TABLE 31

*Reliability Data: Intelligence Scale for Children*[27]

| | | *Subtest* | | |
|---|---|---|---|---|
| *Age Group* | *Range of r's* | *Mean* | *High r* | *Low r* |
| 7½ | .59–.84 | .67 | Block Design | Comprehension and Picture Completion |
| 10½ | .59–.91 | .76 | Vocabulary | Digit Span |
| 13½ | .50–.90 | .75 | Vocabulary | Digit Span |

*IQ Reliabilities*

| | *Verbal* | *Nonverbal* | *Full* |
|---|---|---|---|
| 7½ | .88 | .86 | .92 |
| 10½ | .96 | .89 | .95 |
| 13½ | .96 | .90 | .94 |

(Digit Span, Coding and Mazes are not included.)

individual's "true" IQ on this scale is within 4.25 points of his obtained IQ.

*Conclusions on Reliability.* The reliability coefficients and the standard errors of measurement must be taken into account when scores on the individual subtests are being interpreted or when differences in scores between subtests are being evaluated. The lower the reliability and the larger the standard error, the less is the confidence to be placed in judgments based upon scores of that particular subtest.

Since, on the basis of the standardization data, the reliabilities of the several IQ's are at a satisfactory level, it appears that considerably more confidence can be placed in those indexes than in the scores of the individual subtests (with the exception of vocabulary).

[27] From the *Manual*, p. 13. The Psychological Corporation. (By permission.)

Since there are marked differences between reliability coefficients of the subtests for each of the three age groups reported in the standardization data, it is highly desirable that separate reliability studies be made for each of the eleven age groups separately, especially at the extremes of the age distribution (5 to 15) for which the scale is intended

TABLE 32

Standard Errors of Measurement Intelligence Scale for Children[35]

| | | | Subtest | |
|---|---|---|---|---|
| Age Group | Range * | Mean | High | Low |
| 7½ | 1 20–2 45 | 1 74 | Digit Span | Picture Arrangement |
| 10½ | 90–1 92 | 1 44 | Digit Span | Vocabulary |
| 13½ | 95–2 12 | 1 47 | Digit Span | Vocabulary |

*IQ Standard Errors* **

| | Verbal | Nonverbal | Full |
|---|---|---|---|
| 7½ | 5 19 | 5 61 | 4 25 |
| 10½ | 3 00 | 4 98 | 3 36 |
| 13½ | 3 00 | 4 74 | 3 68 |

* Standard errors of measurement of subtests are given in units of the weighted scores

** Standard errors of intelligence quotients are given of course, in IQ points

**Validity.** *Subtest Intercorrelations* In the manual for this scale, there are no data on the problem of validity as such There are data on intercorrelations of the subtests The assumption is that significant intercorrelations between subtests would validate the hypothesis that they and the scale as a whole measure common factors However, the intercorrelation coefficients among the individual subtests are, on the whole, not as high as would be expected At the 7½-year level, these coefficients are concentrated within the 20's and 30's, at the 10½ year level, they are concentrated within the 30 s and 40's, while at the 13½ year level, they are distributed within the 20's, 30's, and 40's

On the other hand *each verbal subtest* correlates quite significantly with *total verbal score*, the range for the three age groups being from

[35] From the *Manual* p 13, The Psychological Corporation (By permission )

The standard error of measurement indicates the range of score within which the chances are approximately two to one that a subject's "true" score will fall in that particular subtest Thus, the standard error of 1 20 for 7½-year-olds on picture arrangement indicates that the probabilities are two to one that an individual's "true" score on this subtest is within 1 20 points of his obtained weighted score Likewise the standard error of 4 25 IQ points (full scale) for 7½-year-olds indicates that the probabilities are about two to one that an

TABLE 31

*Reliability Data Intelligence Scale for Children*[37]

*Subtest*

| *Age Group* | *Range of r's* | *Mean* | *High r* | *Low r* |
|---|---|---|---|---|
| 7½ | 59–84 | 67 | Block Design | Comprehension and Picture Completion |
| 10½ | 59–91 | 76 | Vocabulary | Digit Span |
| 13½ | 50–90 | 75 | Vocabulary | Digit Span |

*IQ Reliabilities*

| | *Verbal* | *Nonverbal* | *Full* |
|---|---|---|---|
| 7½ | 88 | 86 | 92 |
| 10½ | 96 | 89 | 95 |
| 13½ | 96 | 90 | 94 |

(Digit Span, Coding and Mazes are not included )

individual's "true" IQ on this scale is within 4 25 points of his obtained IQ

*Conclusions on Reliability* The reliability coefficients and the standard errors of measurement must be taken into account when scores on the individual subtests are being interpreted or when differences in scores between subtests are being evaluated The lower the reliability and the larger the standard error, the less is the confidence to be placed in judgments based upon scores of that particular subtest

Since, on the basis of the standardization data, the reliabilities of the several IQ's are at a satisfactory level, it appears that considerably more confidence can be placed in those indexes than in the scores of the individual subtests (with the exception of vocabulary)

[37] From the *Manual* p 13 The Psychological Corporation (By permission )

Since there are marked differences between reliability coefficients of the subtests for each of the three age groups reported in the standardization data it is highly desirable that separate reliability studies be made for each of the eleven age groups separately, especially at the extremes of the age distribution (5 to 15) for which the scale is intended

TABLE 32

*Standard Errors of Measurement Intelligence Scale for Children*[38]

| | | | *Subtest* | |
|---|---|---|---|---|
| Age Group | Range * | Mean | *High* | *Low* |
| 7½ | 1 20–2 45 | 1 74 | Digit Span | Picture Arrangement |
| 10½ | 90–1 92 | 1 44 | Digit Span | Vocabulary |
| 13½ | 95–2 12 | 1 47 | Digit Span | Vocabulary |

*IQ Standard Errors* **

| | Verbal | Nonverbal | Full |
|---|---|---|---|
| 7½ | 5 19 | 5 61 | 4 25 |
| 10½ | 3 00 | 4 98 | 3 36 |
| 13½ | 3 00 | 4 74 | 3 68 |

* Standard errors of measurement of subtests are given in units of the weighted scores

** Standard errors of intelligence quotients are given of course in IQ points

Validity *Subtest Intercorrelations* In the manual for this scale, there are no data on the problem of validity as such There are data on intercorrelations of the subtests The assumption is that significant intercorrelations between subtests would validate the hypothesis that they and the scale as a whole measure common factors However the intercorrelation coefficients among the individual subtests are, on the whole not as high as would be expected At the 7½ year level these coefficients are concentrated within the 20 s and 30 s, at the 10½ year level, they are concentrated within the 30 s and 40 s while at the 13½ year level they are distributed within the 20 s, 30 s and 40 s

On the other hand *each verbal subtest* correlates quite significantly with *total verbal score* the range for the three age groups being from

[38] From the *Manual* p 13 The Psychological Corporation (By permission )

44 to 82, with the coefficients fairly evenly distributed over this range The *nonverbal subtests* correlate somewhat lower with *total performance scores*, the range being from 32 to 68, with some concentration in the 50's

The correlation coefficients between *total verbal scores* and *total performance scores* are, respectively, 60, 68, and 56 for these same age groups

TABLE 33

*Correlations between the Intelligence Scale for Children and Other Scales*
*(5 Studies)*

| *Other Scale* | *Subjects* | *Number* | *r* |
|---|---|---|---|
| Arthur Point Scale | mentally defective | 40 | 79 (Full Scale) |
| " " " | " " | 40 | 83 (Nonverbal Scale) |
| " " | " " | 40 | 47 (Verbal Scale) |
| Stanford Binet, L | " " | 40 | 76 (Full Scale) |
| " | " " | 40 | 64 (Nonverbal Scale) |
| " " | " " | 40 | 75 (Verbal Scale) |
| Stanford Binet | subnormals | 70 | 68 (Full Scale) |
| " " | " | 70 | 69 (Verbal Scale) |
| Stanford Binet | normals | 49–53 | 85 (Full Scale) |
| " " | " | 49–53 | 82 (Verbal Scale) |
| " " | " | 49–53 | 80 (Nonverbal Scale) |
| Arthur Point Scale | " | 49–53 | 80 (Full Scale) |
| " " " | " | 49–53 | 77 (Verbal Scale) |
| " " | " | 49–53 | 81 (Nonverbal Scale) |
| Stanford Binet L | " | 54 | 80 (Full Scale) |
| " " | " | 54 | 71 (Verbal Scale) |
| " " | " | 54 | 63 (Nonverbal Scale) |
| " " | " | 332 | 82 (Full Scale) |
| " " | " | 332 | 74 (Verbal Scale) |
| " " | " | 332 | 64 (Nonverbal Scale) |

These findings indicate that, on the whole, while each subtest has only a very moderate amount of communality with the others taken singly, verbal subtests *combined* have much more communality with each *individual verbal subtest*[29] The *same is true of combined* performance and separate performance scores

Finally, the data indicate that *all the verbal subtests taken as a whole* have considerable communality with *all the performance sub-*

[29] Corrections are made in order to eliminate self-correlation

*tests as a whole* Yet, since the aforementioned coefficients of 60, 68, and 56 are fairly distant from unity, the measured abilities in one group (verbal) can be used only for a general approximation of abilities measured by the other group of subtests (nonverbal), and *vice versa* The reporting therefore, of verbal, nonverbal, and full

TABLE 34

*IQ s of Intelligence Scale for Children Compared with Two Other Scales*
*Means and Standard Deviations*
*(5 Studies)*

| WISC | Arthur Point Scale | S B | Subjects | N |
|---|---|---|---|---|
| 60(S D 6) Full<br>65(S D 13) Verbal<br>58(S D 10) Perform | 65(S D 12) | 56(S D 5) | Deficients | 40 |
| 66(S D 9) Full<br>67(S D 7) Verbal<br>72(S D 11) Perform | | 65(S D 7) | Subnormal | 70 |
| 100(S D 15) Full<br>99(S D 14) Verbal<br>101(S D 15) Perform | 95(S D 16) | 105(S D 15) | Normal | 49–53 |
| 102(S D 11) Full<br>101(S D 12) Verbal<br>104(S D 11) Perform | | 106(S D 11) | Normal | 54 |
| 101(S D 13) Full<br>103(S D 14) Verbal<br>98(S D 15) Perform | | 108(S D 16) | Normal | 332 |

scale IQ s with this instrument is a desirable, in fact a necessary, practice

*Correlations with Other Scales* Since the appearance of this scale several reports have been published that deal with the correlations and IQ differences found between it, the S B and the Arthur The summarized data are given in Tables 33 and 34 Unfortunately, the findings of these studies must, with one exception, be regarded as only suggestive and quite tentative, for the number of cases in each is very small, or the coefficients have been affected by the age range of the testees

The exception reports on 332 cases between the ages of 5 and 15

The data given in Tables 33 and 34 are for the entire group At the different ages, the *r s* between Stanford Binet and full scale IQ s vary from 75 to 90, for the verbal scale, between 65 and 90, for the performance scale, between 50 and 75 The table giving mean intelligence quotients and standard deviations indicates that the Wechsler scale tends to rate subnormal subjects somewhat but not markedly higher than does the Stanford Binet At the average level the reverse is true The differences between the means are fairly marked in the study of 332 individuals This being the most comprehensive and detailed report of those herein reported its findings carry the *greatest* weight [40]

TABLE 35

*Correlations of Intelligence Scale for Children with School Achievement*

| *Scale* | *N* | *Range of r s for Separate Subjects* | *Total Achievement Test* |
|---|---|---|---|
| Full | 54 | 45–71 | 76 |
| Verbal | 54 | 48–60 | 62 |
| Nonverbal | 54 | 41–64 | 65 |
| Full | 18–21 | 44–81 | – |
| Verbal | 18–21 | 47–74 | – |
| Nonverbal | 18–21 | 29–74 | – |

On the basis of the research thus far reported, it is reasonable to conclude that full scale intelligence quotients and verbal scale intelligence quotients, on the one hand, and Stanford Binet IQ s on the other, have considerable communality of psychological functions being measured The performance scale intelligence quotients have much less in common with the Stanford Binet

*Predictive Efficiency* Validity of a test for children should also be evaluated in terms of its predictive efficiency with respect to educational achievement In this area too few data are available for this children's intelligence test Table 35 summarizes the results reported in two studies

Correlations of IQ with teachers' ratings of their pupils' intelligence were 68 (full), 64 (verbal), and 53 (performance)

[40] J I Krugman "Pupil Functioning on the Stanford Binet and the Wechsler Intelligence Scale for Children" *Journal of Consulting Psychology* Vol 15 pp 475–483 1951

*Conclusions on Validity* On the whole, these results are encouraging, the correlation coefficients fall within the approximate range of indexes usually found for other widely used tests of intelligence, including the Stanford-Binet These and similar data are, however, not yet definitive, the number of cases is small in both studies, and in one of the studies, grade level and age range were not sufficiently controlled Further research is necessary at each of the age and grade levels for which the scale is intended before generalizations may be offered with assurance regarding the scale's predictive efficiency in respect to educational achievement

**Evaluation and Criticisms.** On the whole, this intelligence scale for children is a useful addition to the very limited number of instruments now available for individual testing It is to be expected that the deficiencies regarding standardization population and studies of reliability and validity will be remedied as the scale continues to be used for both practical and research purposes Some psychologists, at this stage, question the wisdom of using this Wechsler scale as a substitute for the Stanford-Binet until more definitive data are obtained for the former

Although one of the advantages originally claimed for this scale was that it did not use the mental age concept, it has since been found desirable to supply mental-age equivalents [41] The mental age concept is an extremely useful one when interpreted by qualified psychologists This concept should be made an integral part of the scale

The relatively low reliabilities of the *subtests* indicate that there is no merit in merely deriving a test profile for purposes of diagnosis and guidance The reliabilities of part-scores must be high before profiles can be used with confidence The *total* verbal, performance, and full scores, however, have yielded reliability coefficients at a satisfactorily high level of confidence

Considerably more research remains to be done on the predictive efficiency (validity) of this scale Data available thus far show that IQ differences between it and the Stanford Binet are significant enough to warrant caution, in spite of the generally high correlations found between them, and, it must be added, some of the coefficients

---

[41] D Wechsler, "Equivalent Test and Mental Ages for the WISC," *Journal of Consulting Psychology*, Vol 15, pp 381–384, 1951

between the two scales are only moderate.[42] The Stanford-Binet IQ's tend to be higher within the "normal" range, especially at the earlier age levels Since the Stanford-Binet has been in use much longer, and has been found to have considerable value in schools and clinics, and has been widely used as a validating criterion, it is probable the obtained discrepancies between the two scales will stimulate research on and improvement of the more recent instrument

The limits of the IQ values given by the Wechsler full scale are from 46 to 154 This means that the scale cannot be used with individuals who rank above or below these limits In terms of total number of persons, the percentages of such cases will be very small, but in particular instances this can be a serious limitation

[42] J J Pastovic and G M Guthrie, "Some Evidence on the Validity of WISC," *Journal of Consulting Psychology* Vol 15, pp 385-386, 1951 A particularly significant finding in this report is that, on the Wechsler scale, the mean performance IQ s are higher than the mean verbal IQ s If these results are corroborated they will raise some doubt over the applicability of the scale to the usual problems of predicting and evaluating educational achievement.

# 8.

## INDIVIDUAL PERFORMANCE SCALES

ABOUT the time the final revision of Binet's own scale appeared, some psychologists in the United States had assembled a group of performance tests intended to meet practical problems in the study of human abilities and behavior One of these was the Healy-Fernald group of tests, devised primarily to examine juvenile delinquents in an effort to determine their intellectual levels and personality traits revealed in the course of the examination[1] Unlike tests which were subsequently developed and which are now in use, those of Healy and Fernald were not actually standardized in respect to administration and scoring This group of tests provided the psychological examiner with situations wherein he could observe, evaluate, and interpret the testee's methods of solving problems and his behavior in test situations The specific tests were selected on the basis of Healy's and Fernald's judgment and psychological insights as to what constitutes intelligent activity, beyond this the value of the results obtained with their test would depend upon the clinical acumen of examiners, since there were no norms based upon standardization procedures While the Healy-Fernald tests are infrequently used today, they are important in the historical development of performance scales

Several of the older and of the current scales will be described, these descriptions will be followed by a discussion of their uses and by a general evaluation of this type of instrument

[1] W Healy and G M Fernald *Tests for Practical Mental Classification*, Psychological Monographs Vol 13, No 2 1911

## THE PINTNER-PATERSON SCALE OF PERFORMANCE TESTS

*Contents.* This group of performance tests, the first to be organized into a scale, is now of interest principally for its historical and background value.[2] It will also enable the student to see that many of the earliest types of performance tests have survived the years of experimentation and application and have been incorporated into scales now current, including the Bellevue.

Pintner and Paterson standardized some of the Healy-Fernald performance tests as well as several which had been devised by other psychologists and themselves. The final scale includes fifteen tests which can be presented without the use of language, nor do they require the use of language on the part of the subject. They are intended primarily for use with persons having serious hearing defects and for non English speaking individuals. These and similar performance tests have been found valuable as supplements to verbal tests of mental ability, and also with subjects who, though they are English-speaking, have speech defects or reading disabilities.

The subtests in the scale are described below.

(1) *Mare and Foal Form Board.* This, of the picture puzzle type, is a pictureboard of a mare and foal in color. Sections of the board are removed to begin with; the subject must replace them correctly. Score is based on time required and number of wrong moves.

(2) Seguin Form Board. This is a form board in which ten common geometric shapes are to be placed. Score is based on the shortest time required in three trials.

(3) Five Figure Board. There are five geometric figures, each of which is divided into two or three parts. The pieces are to be fitted into their appropriate places. Score is based on time required and number of errors made.

(4) *Two-Figure Board.* There are two geometric figures, one cut into four sections and the other into five. These are to be correctly placed in two spaces. Score is based on time required and number of moves.

(5) Casuist Board. This form board—more difficult than the preceding ones—consists of four spaces in which twelve sections have to be fitted. Score is based on time required and number of errors made.

---

[2] R. Pintner and D. Paterson. *A Point Scale of Performance Tests.* New York: D. Appleton, 1917.

(6) Triangle Test Four triangular pieces are to be fitted into the board Score is based on time required and number of errors made

(7) Diagonal Test Five variously shaped sections have to be fitted into a rectangular form Score is based on time required and number of errors made

(8) Healy Puzzle A This consists of five rectangular sections which are to be fitted into a rectangular frame Score is based on time required and number of moves made

(9) Manikin Test Wooden legs arms head and body are to be put together to make the form of a man Score depends on quality of performance

(10) Feature Profile Test Wooden sections have to be put together to form the profile of a man s head Score is based on time required

(11) Ship Test (originated by H A Knox) This is a picture of a ship cut into ten sections all of same size and shape to be inserted properly in a rectangular frame Score depends on quality of performance

(12) Healy Picture Completion Test I This is a large picture from wh ch ten small squares have been cut out The missing parts are to be selected from among forty-eight squares identical in size Score depends on the quality of completion within a limit of ten minutes

(13) Substitution Test A page of rows of geometric figures (five different shapes) which have to be marked with appropriate digits to correspond with a key at top of page Score is a combination of time and errors made

(14) Adaptation Board This is a form board having four circular blocks and holes three are 6 8 cm in diameter while the fourth is 7 cm The subject is shown that one block fits the larger hole He is then required to keep his attention fixed and to fit this larger block into the correct space when the board is moved into four different positions Score is based on the number of correct moves

(15) Cube Test Four cubes (one inch) are placed before the subject With a fifth cube they are tapped in a specified order by the examiner The subject is asked to imitate the order of tapping The sequence becomes longer and more complex Score is the number of sequences correctly imitated

For general testing purposes the authors of this performance scale recommend the use of a short scale which includes ten of the fifteen parts namely, 1 2, 3 4 5 9 10 11, 12 and 15 of the foregoing list

The age range of the Pintner Paterson scale is from four years to fifteen However this does not mean that every test in the series has

discriminative value throughout this range For example, the Seguin Form Board does not have value in general beyond age ten, while the Feature Profile test is not generally useful below age ten

**Scoring.** Three different methods of scoring were provided by the authors median mental age, point score, and percentile rank For each test there is a separate table of mental age norms, the median of an individual's MA's on each of the several subtests is taken as the single MA to represent his general performance on the entire scale In the point scale, the subject earns a total score in points for all the parts, the total score determines his MA, as indicated in a table of norms By the percentile method, the subject's score on each of the several tests yields a percentile score, these can be combined to yield a single percentile rating for the whole scale Of the three indexes, the median mental age has been used most widely with this scale

**Evaluation.** During the many years that the Pintner-Paterson performance tests were used for clinical and experimental purposes, the following evaluations of them were widely accepted They are more susceptible to practice effects, and chance successes are more frequent, than is the case with verbal tests, hence, the reliability coefficients of these performance tests are not as high as those of the verbal This scale is useful primarily with young children, and with older children and adults who are mentally retarded or deficient The scale has clinical significance, also, in the case of an older child when there is marked discrepancy in performance on the several subtests The several parts of the performance scale examine processes that are more specific than those examined by verbal tests This is indicated by the considerable scatter of ratings on the several parts and by the lower correlation coefficients found when ratings on each separate part were correlated with ratings for the whole scale, the range being from very negligible coefficients to fairly high, with a median at about 50

Performance tests of the Pintner-Paterson kind correlate poorly or only very moderately with intelligence tests of the verbal kind as represented by the Stanford Binet, when the group being studied is limited with respect to age or range of ability For instance,[3] when a group of gifted children was examined with both scales, the correla-

[3] D A MacMurray, "A Comparison of Gifted Children and of Dull Normal Children Measured by the Pintner Paterson Scale as Against the Stanford Binet Scale," *Journal of Psychology* Vol 4 1937, pp 273-280

tion coefficient between the two sets of ratings was only 23 For a group of dull children who were likewise examined the coefficient was 43 Assuming adequate reliability of both scales these coefficients indicate that there is only little or moderate correspondence between them regard ng functions being tested Furthermore the very low coefficient of 23 suggests also that the performance scale

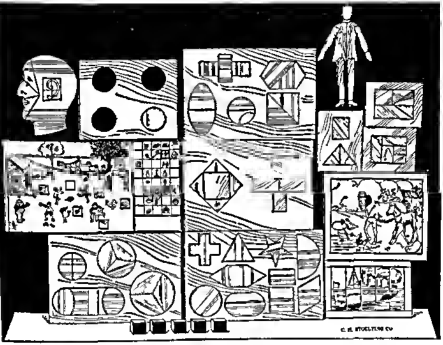


FIG 8 1 Pintner Paterson Performance Tests C H Stoelting Company (By permiss on )

is particularly inadequate for differentiating among performance levels of gifted children

A fairly large number of studies have been published reporting higher correlations between performance test ratings obtained with the Pintner Paterson and similar tests and those obtained with re visions of the Binet scale Coefficients of the order of 70 and 80 were not uncommon whereas others were as low as 50 These co efficients however cannot be interpreted as necessarily indicating that there is a considerable community of function between these performance tests on the one hand and verbal tests of ability on the

other Indeed, it appears that to a considerable degree the correlation coefficients are due to the wide age-range of the subjects tested, with the result that the coefficients reflect the fact that the psychological functions being tested by both types increase with age, that is, the results on both types of tests are to an appreciable extent the product of age An ordinary group of ten year-old children will get higher scores on both types of tests than will a similar group of nine-year-olds, who, in turn, will score higher on both than an ordinary group of eight-year-olds, and so on This is to be expected, for the tests have been so constructed as to yield progressive increases in age-norms as chronological age increases

Another example of the effect of age-range on correlations is found in the coefficients between intelligence ratings and height or weight or dentition These, for a wide age range, are in the neighborhood of 50 and 60, because in general, older children are taller, heavier, and have more permanent teeth It would not be said, of course, that these coefficients indicate community of function between the psychological tests and the physical measures But within a single age group the correlation coefficients between these physical traits and intelligence test ratings drop down to negligible levels Thus, when age is held constant, or very nearly so, the correlation coefficients between results obtained with the Pintner-Paterson and similar performance tests, on the one hand, and verbal tests, on the other, drop to between 40 and 60

A factorial analysis of results obtained on thirty four commonly used performance tests suggests one reason why there is only a low or moderate correlation between these and verbal tests of mental ability This analysis appears to indicate that the principal factors measured by the performance tests may be identified as "spatial, perceptual speed, and induction "[4] While the first two of these functions are involved to some extent in many verbal tests of ability, they are actually of relatively little significance there in the determination of an individual s rating.

These findings signify that verbal scales and those of the Pintner-

[4] C M Morris, "A Critical Analysis of Certain Performance Tests" *Pedagogical Seminary and Journal of Genetic Psychology* Vol 54 1939 pp 85–105 "Perceptual speed" is the readiness to discover and identify perceptual detail (mainly visual) The "spatial" factor involves the ability to manipulate objects in space

Paterson type may not be used interchangeably but should be used to supplement each other

## THE CORNELL-COXE PERFORMANCE ABILITY SCALE[5]

**Contents.** For this scale, the particular tests included were selected from a variety of sources The authors carried out their own standardization and revised the directions for administering and scoring

The tests included are the following

(1) Manikin and Profile These are already familiar to the reader They are scored for accuracy and time required

(2) Block Designs These are the familiar Kohs colored block designs five of which were included They are scored for accuracy and time required

(3) Picture Arrangement This includes ten series of pictures which though different in subject matter, are the same in principle as those in the Bellevue scale They are scored for accuracy only

(4) Digit Symbol This type test is also familiar to the reader for it is of the same kind as that included in the Bellevue and other scales It is scored for accuracy and time required

(5) Memory for Designs This test includes five cards on each of which is a geometric design The subject is asked to reproduce each design after it has been shown for ten seconds This type of test is similar to that used by Binet The score depends upon quality of reproduction

(6) Cube Construction This test utilizes blocks some sides of which are painted while others are not The examiner presents models of cube construction and asks the subject to duplicate them The score depends upon both accuracy and time

(7) Picture Completion (This is an optional substitute for test 3 ) The Healy Picture Completion Test II was selected The score depends upon accuracy only (This test is the same in principle as Picture Completion I, but its theme is different and on a higher level of difficulty )

*It will be noted* that the Cornell Coxe scale *differs from other performance* scales in that it does not include any form boards

Scoring In order that each test in the scale might contribute equally to the total score, the authors follow the common practice of converting raw scores into a type of standard unit[6] Total score is ob-

[5] E L Cornell and W W Coxe *A Performance Ability Scale* Yonkers N Y World Book 1934

[6] *Ibid* pp 29 ff

tained by adding the weighted scores for the several parts A mental age is then obtained from a table of norms, extending from an MA of 4 years and 6 months to 16 years and 8 months

**Validity and Reliability.** It appears from experimental data, cited by the authors, that they used the following criteria of validity a "*satisfactory*" *distribution of scores by* school grades (*increasing* averages in successive grades), a distribution of scores that conforms well to the symmetrical bell-shaped curve, a high correlation between scores on each test in the scale with total scores Although the correlation between total performance scores and chronological ages is .78, Cornell and Coxe do not regard CA as a validating criterion of first importance

Reliability coefficients of the several parts of the scale varied from .66 to 89, while *for total scores the reliability coefficient* was 929 (125 cases) These data indicate a satisfactory degree of reliability for the scale as a whole

The authors of this scale *maintain that* a performance scale should not be a substitute for those of the Binet type and other verbal tests, but should supplement them They sought, therefore, to devise an instrument which should differ from these others in respect to functions tested In this, they apparently succeeded fairly well, for although the correlation coefficient found between total performance score and Stanford-Binet (1916) mental ages was .79 for a wide age-range, when chronological age was held constant the partial correlation coefficient was reduced to 38 This means that in a group of about the same age the results of the two scales would indicate relatively low community of function [7]

It appears also that the several tests within the scale have only very moderate community of function Intercorrelations of the parts ranged from coefficients of 50 to 75, over the entire age range, but when chronological age was held constant, the partial correlation coefficient varied from about 20 to 60

In constructing their scale, Cornell and Coxe were interested primarily in developing a supplementary instrument They say in this regard "One important value of any scale supplementary to the

[7] Practically the same results were obtained when the Cornell-Coxe scale was correlated with the National Intelligence Test a group test of the verbal type The *simple correlation coefficient* was found to be 74

Binet scale lies in the fact that if the two scales used give different results, the psychologist's attention is directed toward discovering reasons for whatever differences may be found, and his analysis and interpretations are thereby enriched and tend to have greater validity" [8] In view of the data on the degree of correspondence between results obtained with this performance scale, on the one hand, and those obtained with Binet revisions and verbal group tests, on the other, it is reasonable to conclude that the Cornell Coxe scale serves the purpose for which it is intended

## THE ARTHUR POINT SCALE OF PERFORMANCE TESTS [9]

**Contents.** Form I of this scale is a restandardization of some of the tests used in the Pintner-Paterson, plus two other tests The eight parts are Knox Cube Test, Seguin Form Board, Two Figure Form Board, Casuist Form Board, Manikin, Feature Profile, Mare

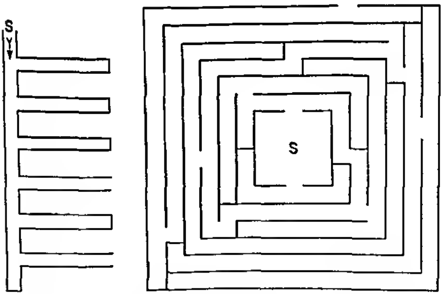



FIG 8 2 Porteus Maze Tests—*Years 5 and 14* C H Stoelting Company (By permission)

---

[8] *Ibid* p 37

[9] G Arthur, *A Point Scale of Performance Tests* New York The Commonwealth Fund, Vol 1, 1930, Vol 2, 1933, Vol 1 appeared, also, in a revised edition in 1943

and Foal Healy Picture Completion I The two additional are Porteus Maze Test and Kohs Block Design Test

The Porteus test consists of a series of mazes of increasing difficulty each printed on a separate sheet The subject is required to trace with pencil the course from entrance to exit The Kohs test consists of the same set of blocks used in the Bellevue scale but different designs are to be reproduced

The purpose of the restandardization of the scale is to provide a more reliable and useful performance scale for clinicians Form I is based upon results obtained with 1100 public school children of middle-class families The usual validating criteria were applied such as parental occupation age grade distribution significant increases in score in successive ages and degree of correspondence with ratings obtained by means of other scales already considered to be acceptably valid (Stanford Binet and Kuhlmann Binet)

Scoring An individual s score is variously determined on the several tests by number of successes or time required or degree of accuracy or a combination of these Each test yields a raw score which is converted into weighted score points[10] The raw score for each subtest is assigned a value proportional to the effectiveness of the subtest in differentiating between successive age levels. The total of these weighted scores is converted into a mental age The range of mental age norms in Form I is from five and a half years to fifteen and a

[10] The formula assumes that the value of a test, hence its weight in the total score depends upon the extent to which it differentiates between successive age groups The greater the difference between age-group averages of a test and the less the amount of overlapping of scores of two adjacent age groups, the greater is the weight given to that test. The formula is

$$DV \quad \frac{M - M}{\frac{PE + PE_y}{2}}$$

in which $M$ is the mean of the older of the two adjacent age groups $M$ is the mean of the younger age group the *PE*'s are the probable errors (middle 50 percent of the scores) of the two groups *DV* stands for "discriminative value" Inspection of this formula shows that as the difference between means increases, the numerator grows larger and discriminative value increases But the *DV* is also dependent upon the sizes of the probable errors which indicate the amount of overlapping of scores between the age group Thus also as the probable errors are smaller and therefore the overlapping less the denominator is the smaller and the fraction—hence the *DV*—is larger See Arthur *op cit* Vol 1 1943 p 39

half Arthur employs a statistical device to extend the norms downward by six months by using a constant monthly rate of decrease in score This is a procedure of very doubtful validity because it assumes that rate of psychological development—as represented by these tests —is constant in the early years, whereas the prevailing conception among psychologists is that rate of development is most rapid in the earliest years and decreases as the child grows older [11] The scale is similarly extended at the upper end by adding a constant value, in points, to provide hypothetical mental age ratings beyond the norms derived by the actual standardization process [12] When mental ages and intelligence quotients are obtained from norms thus extrapolated, the examiner must clearly realize that he is deriving indexes which do not necessarily have the same meaning as those found by actual standardization

## REVISED ARTHUR SCALE FORM II

The purpose of the second form is to serve as an alternate when retesting and for use with preschool children

This version of the scale utilizes four of the test types already described, namely, Knox Cube, Seguin Form Board Porteus Mazes, and Healy Pictorial Completion II The only new type of material not thus far described is the Arthur Stencil Design This test employs twenty designs increasingly complex and more difficult to reproduce, which are presented singly The testee is given six square colored cards and twelve colored stencils which are cut within square cards Each design is to be reproduced by placing the appropriate cards and stencils, one upon another, so as to duplicate the original in both form and color For example, a practice design requires merely that a red octagonal stencil be laid over a white card to get the desired result

**Reliability.** Scores of Forms I and II were correlated at each age level from six to sixteen years The coefficients, with CA constant, ranged from 55 (PE ± 06) at age 8, to 70 (PE ± 05) at ages 10

[11] Dr Arthur herself recognizes this

[12] The reader should contrast this with the method employed in constructing the Bellevue scale remembering however that the Bellevue is designed primarily for adults while the Arthur scale is not Compare also with method of deriving mental ages and intelligence quotients by means of the Stanford Binet scale at levels of superior adults

and 15 The median coefficient was 61 (PE ± 06) As estimates of reliability, these coefficients are relatively low The small number of cases in each age group, varying from 41 to 54, might account in part for these results

That the scale is more reliable than the foregoing data suggest is indicated by the results of another study in which the subjects were 61 institutionalized mentally deficient boys whose mean IQ on the Stanford Binet was 67 They were tested with the Arthur scale then retested after an interval of two years, the correlation between the two sets of scores being 85 The coefficients for each of the parts varied from 69 to 80 The over all results indicate satisfactory stability of relative rank, especially in view of the narrow range of ability within the group of subjects tested It is to be noted however that the mean gain in the Arthur scale IQ was ten points as contrasted with a mean loss of only one IQ point on the Stanford Binet during the same interval [11]

The mean gain of ten points may be attributed to one or both of two factors (1) scores on the performance type of test are more susceptible to practice (learning) than are those on the verbal type, or (2) residence and training in a soundly conceived and operated institution encourages the development and utilization of the *potentialities* of the mentally deficient beyond levels attained under ordinary circumstances The latter factor is not synonymous with specific practice and learning effect It is rather, a result of general training in more effective behavior, and, in some instances, also the removal of "blocks" that impair one s performance

**Validity.** This performance scale was devised primarily as a clinical instrument to be used as a substitute for the Binet revisions in cases where a verbal type of scale is inappropriate as in instances of language handicap defects of vision or hearing and inequality in development of an individual s verbal and nonverbal functions Arthur in her standardization procedures has taken the position that the basic capacities demanded by the Binet tests and by her performance tests should be essentially the same thus the Kuhlmann and Stanford revisions of the Binet are two principal validating criteria

[11] R. M. Patterson, "The Significance of Practice Effect upon Readministration of the Grace Arthur Performance Scale to High Grade Mentally Deficient Children," *American Journal of Mental Deficiency* Vol 50 1946 pp 393-401

used in constructing her scale Accordingly, the main differences between the Arthur and the Binet revisions should be in the types of materials used to sample the psychological functions

The extent to which the Arthur and Binet scales actually correspond may be inferred from a comparison of IQ ratings obtained by subjects examined with both, and from the correlation coefficients obtained between the two sets of ratings, taken separately for each age group

In the first place, Arthur reports that the probable error (PE) of intelligence quotients was 4 97 points when the performance scale

TABLE 36

*Correlations of Stanford Binet IQ's and Arthur IQ s*

| Age | N | R |
|---|---|---|
| 5 | 35 | 70 ± 06 |
| 6 | 54 | 77 ± 04 |
| 7 | 50 | 68 ± 05 |
| 8 | 44 | 74 ± 05 |
| 9 | 41 | 80 ± 04 |
| 10 | 40 | 51 ± 08 |
| 11 | 44 | 68 ± 05 |
| 12 | 31 | 80 ± 04 |
| 13 | 27 | 21 ± 12 |
| 14 | 27 | 07 ± 13 |
| 15 | 16 | − 10 ± 17 |

ratings were compared with those of the Kuhlmann Binet, and 4 92 points when compared with the Stanford Binet What this means is that in fifty percent of the cases, the IQ differences were five points or less, while in the remaining fifty percent the differences were greater The frequencies of the differences, however, decline markedly as the size of the differences increases beyond five points

Correlation coefficients between Stanford-Binet (1916) IQ's and Arthur scale IQ's (Form I) are shown in Table 36 [14] Most of these coefficients are rather high and noteworthy Excepting at age ten, they show unusually high correspondence for these two types of scales between ages five and twelve The correlations at the later ages, however, are so low and the probable errors (PE) so large that the coefficients

[14] Calculated from data in Arthur, *op cit* Vol 2, pp 54–61

*may be regarded as being zero for all practical purposes* If the table of coefficients is representative of the correspondence existing between Stanford Binet IQ s and Arthur Performance IQ s at ages above twelve, then we must conclude that the latter scale has been inadequately standardized or is incapable of differentiating among individuals at the later age levels in respect to the functions being measured

Although the coefficients for ages five to twelve are quite marked and in several instances high they are, nevertheless not close enough to unity (+1 00) to warrant the use of the Stanford Binet and the Arthur scales interchangeably For clinical purposes, the Arthur scale is valuable within the age range of five to twelve as a supplement to verbal scales of the Stanford Binet type

This conclusion is further supported by validating data obtained subsequent to the publication of Arthur's manual These later studies using both the 1916 and 1937 revisions of the Stanford Binet, can be summarized as follows

Stanford Binet and Arthur IQ s correlate variously, from about 50 to about 80

In a large majority of cases the Arthur scale IQ s tend to be somewhat higher than the S B at levels below 90 IQ

At the levels above 90 IQ the S B tends to yield somewhat higher ratings

The means of the differences between the IQ s of the two scales have been found to range from about 5 to 10 points

While 'discriminative value regarded by Arthur as a most important criterion of validity is no doubt significant it does not in itself demonstrate that a scale is measuring the functions it has set out to measure

There is therefore need for more definitive studies of the validity of the Arthur scale, using large enough numbers of subjects including not only clinical and institutional groups, but a normal population sampling as well

Arthur s own position appears to be that the appreciable extent of agreement between Stanford Binet test results and those of her own performance scale indicates rather even development and manifestation of psychological functions in general But she believes if

the results of these two scales disagree significantly in the case of a given individual, this is due to unevenness in development and expression of functions, or some complicating nonintellectual factors are responsible for the discrepancy

In the case of any individual instance, the actual interpretation of performance test results, taken in conjunction with verbal test findings, will depend upon all information available with regard to the person concerned and upon the psychologist's interrelating of all relevant facts and data

## OTHER PERFORMANCE TESTS

It is not our purpose to present a description of all available tests of the performance type We have presented several in some detail in order to acquaint the reader with their nature and their uses Those scales described are typical There are, however, several others which are intended to serve a special purpose, of these, three will be very briefly described

**The Ferguson Form Boards**[15] The first description of these was published in 1920 They consist of a series of six form boards used as a unit and progressing in difficulty by fairly equal intervals The tests were standardized upon 364 subjects ranging from children in grade one to college seniors

Ferguson, apparently, used grade placement and school achievement as the principal evaluating criteria, for he reported correlations of his form board scores as follows with grade placement, 81, with teachers' estimates of intelligence 50, with class standing, 56

Since their appearance in 1920, these form boards have been subjected to experimentation from time to time for purposes of revising procedure in administering and scoring, and providing more adequate norms One of the most thorough revisions is that by Wood and Kumin (see footnote 15) who give norms for the ages of 7 years and 6 months to 17 years and 5 months But, as is the case with many other tests of this type, a very large percentage of the standardization population consisted of individuals who had come to a guidance

[15] G O Ferguson A Series of Formboards " *Journal of Experimental Psychology* Vol 2 1920 pp 47–58 L. Wood and E Kumin "A New Standardization of the Ferguson Formboards," *The Journal of Genetic Psychology*, Vol 54, 1939, pp 265–284

clinic for assistance and who, therefore, may not be representative of the general population of their age groups It is necessary to consider this fact when using and interpreting results of performance tests so standardized

For present purposes, the fact of major interest is the extent to which these form boards and the scales of the Stanford Binet type do or do not test common functions Wood and Kumin report the following correlation coefficients Ferguson score and Stanford Binet (1916) mental age, simple correlations, .54 for boys and .55 for girls, partial correlations, when chronological age is held constant, .34 for boys and .47 for girls These coefficients indicate a relatively low community of function between the two tests [16] In this respect, the Ferguson form boards are quite similar to nearly all the other performance tests presented in this chapter

**Kent-Shakow Form Board Series** [17] This series of four form boards devised by Kent and Shakow, was made available in 1925, as the Worcester Formboard Series The series presents eight separate tasks graded in difficulty In 1928 a modified series appeared, available in two forms the *Industrial Model* and the *Clinical Model* which differ in respect to size, the former being the larger The scale was first developed as a clinical instrument, with special reference to the needs of the Worcester (Massachusetts) State Hospital out patient department The standardization population consisted of 150 subjects from the age of six years upward, including adults The early standardization, however, was very inadequate Since these form boards were used frequently with adult subjects, better-standardized adult norms were published in 1939, based upon a population of average and superior adults numbering 355 The authors of this performance test do not attempt to validate it in terms of other established tests Apparently, they are concerned mainly with providing a clinical device

[16] The results were of the same general order of magnitude when the Kuhlmann Anderson tests were correlated with Ferguson scores, namely simple correlation coefficients of .43 for boys and .45 for girls

[17] D Shakow and G H Kent "The Worcester Formboard Series" *Pedagogical Seminary and Journal of Genetic Psychology* Vol 32 1925 pp 499–611 G H Kent and D Shakow "Graded Series of Formboards" *Personnel Journal* Vol 7 1928 pp 115–120 D Shakow and B Pazeian "Adult Norms for the K S Clinical Formboards" *Journal of Applied Psychology* Vol 23 1939 pp 495 502 W R Grove "Modification of the Kent Shakow Formboard Series," *Journal of Psychology* Vol 7, 1939 pp 385–397

which, they believe, measures manipulative skill and form analysis, and which, presumably, also provides a means of observing the subject's modes of approaching a problem

In 1939 Grove published a modification of the Kent-Shakow series, Industrial Model, based upon an experimental group of 300 "native

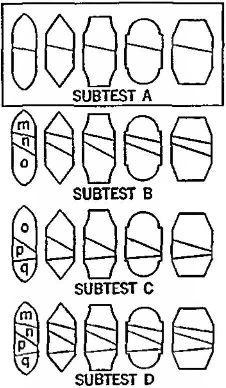



FIG 8 3 *Modified Kent-Shakow Form Board Series* (By permission of William R Grove)

born white adult male prisoners incarcerated in Western Penitentiary (Pennsylvania)" It is especially noteworthy that the scores obtained on this modified series of form boards, when correlated with Stanford-Binet (1916) ratings, yield a coefficient of only 43 ± 03 (PE) The effect of age on this correlation is negligible, since it is based on a comparatively homogeneous group of about 300 adults Thus, there is relatively little community of function measured by these two tests Grove believes that his revised series measures ability to solve prob-

lems presented in the form of concrete spatial relations It is reasonable to assume that the original Kent-Shakow tests measure the same functions as does the modified series, even though the technical terms used to name these functions are different

**The Carl Hollow Square Scale.** This is a form-board test designed for use primarily with adults, though it is usable also with children over ten years of age The test consists of a " . wooden panel in which is cut a 4½ inch square hole, and 29 blocks of varying straight line geometric forms, each having both straight and beveled edges There are long rectangles, short rectangles, three classes of right triangles, diagonally truncated long rectangles, diagonally truncated short rectangles, and overlapping rectangle triangles "[14] The problem for the testee is to fill the hole with sets of blocks, in a series of twenty tasks which become progressively more complex and difficult Performance is scored on the basis of time required and number of moves made Total scores may be converted into IQ's and MA's, or into percentile ranks

This form-board test, its constructor believes, measures the following psychological processes auditory memory (remembering principles and rules given in the instructions), visual memory (recall of partially repetitive patterns), observation and attention to detail, visual imagery involving synthesis and analysis (planning the placement of blocks without actual manipulation), learning (carry over from earlier tasks to subsequent ones)

It appears that the principal validating criterion employed was correspondence with results obtained with other tests having rather wide acceptance (Stanford-Binet, Kohs Block Design, Otis group test, Terman group test, Thorndike-McCall Reading Test) Correlation coefficients between these criteria and the form-board results varied, in the case of adults, from 50 to 80, while in the case of children over 10 years, the coefficients were from about 60 to 80 With adults, the factor of chronological age would not tend to produce spuriously high coefficients, but when the coefficients found with children are being interpreted, due regard should be given to the age range from 10 to 16 years (the theoretical beginning of adult level)

[14] G P Carl "A New Performance Test for Adults and Older Children The Carl Hollow Square Scale" *Journal of Psychology* Vol 7, 1939 pp 179 199

A coefficient of reliability of 87 is reported for this test, indicating an acceptable degree of consistency that compares favorably with other tests of the verbal as well as the nonverbal type

While the author grants that his performance test measures mental abilities which are involved more in the concrete and practical aspects of activity than in the abstract, he also maintains that it is more a measure of general than special ability His view, presumably is based on the correlations with the other tests of general ability already mentioned This question will be dealt with in the section evaluating performance scales as a group

## FUNCTIONS TESTED BY PERFORMANCE SCALES

This discussion will supplement the analysis that was presented in connection with the nonverbal parts of the Stanford Binet and the Wechsler scales The reader should refer to those for more detail

Since all performance tests involve visual perception and manipulation of objects the number of types of items is relatively limited It is not surprising therefore, to find that the range of psychological functions is also restricted This is one reason why the correlations between performance scales and scales of the Stanford Binet type are not higher than they are, since the latter can sample a much wider range of functioning

If the reader will re-examine the descriptions of the fifteen subtests in the Pintner Paterson scale and the few other types introduced in later scales it will be readily apparent that they may all be classified in one of a few categories

geometric form boards, with variations, from the very simple to rather complex

picture form boards (also known as picture completion) of various degrees of complexity

block designs from simple to complex mazes of varying degrees of complexity

recall of geometric designs

picture arrangement

block building

cube sequences (imitating the order of tapping a series of cubes)

digit symbol

With the exception of the last two types, the performance items test visual perception plus more or less visual insight requiring analysis and synthesis Performance on all the eight types also reflects motor speed in varying degrees, and performance on these is facilitated by visual imagery, that is, by the ability to analyze or synthesize a pattern imaginally before actually going through the movements

Block building and cube tapping sequences are largely matters of imitation requiring the functioning of visual imagery and recall The digit symbol test, as already indicated in connection with the Bellevue, utilizes immediate rote recall and visual imagery

In all of the performance tests, visual motor integration, affecting the speed with which a person responds, is involved

Also, many clinical psychologists hold that performance tests provide an estimate of the subject s attention span, especially in the case of mentally retarded and deficient individuals Attention span, however, is not a process in the sense that visual analysis, memory span, etc , are Attention is rather, an *attribute* of the situation in which an individual is placed If the testee is interested in the task at hand, and if the test is within the range of his apprehension, he will be attentive If he does not understand the task and if he is unable to make any progress with it and is confronted by repeated failure, he will very probably be inattentive

It will be noted that these test items make few or no demands upon abstraction [19] concept formation, or the necessity of transcending the immediate concrete situation For this reason, performance tests are regarded as having limited value as measures of general capacity, especially in the testing of individuals who are above average level

## EVALUATION OF PERFORMANCE TESTS

These tests were first constructed as *substitutes* for verbal scales of the Binet type Many correlational studies showed, however, that it is sounder practice to regard the former as *supplements* to the latter The reason for this interpretation is that, when allowance is made for the factor of chronological age, psychologists have found in almost every instance that the coefficients of correlation fall at 50 or lower—generally lower Hence, although the two types of tests measure some functions in common, or are in other ways interrelated,

[19] "Abstraction" means the separation of a quality or an idea, or a principle from the organization and details of a concrete situation

each type also measures functions different from those of the other type

Performance tests have been found most useful with persons handicapped by language disabilities the deaf, the non English speaking groups the illiterate, and those who have speech or reading difficulties

These tests are valuable also in helping to identify children who are shy or inarticulate because of emotional reasons and who, therefore, may appear at a disadvantage on verbal tests of mental ability

Performance tests, used together with the verbal type, are helpful in identifying the mentally deficient and mentally retarded with increased certainty In cases involving diagnosis of mental deficiency, it is often desirable to supplement the Stanford Binet or a similar scale, *with performance tests in order to check on the probable roles* of the language factor lack of cultural opportunities and poor educational experience in order to estimate to what extent these might have adversely affected the testee's score on the verbal type of scale If a significant difference is found between the two obtained ratings, further study of the individual is indicated before a conclusion is reached It has often been found that mentally deficient and mentally retarded persons obtain somewhat higher ratings on performance than on the Stanford Binet and similar scales The differences however, are not always significant enough to raise a question regarding the diagnosis Furthermore it is to be expected that in many cases tested, the differences would be in the direction stated since the performance tests were given for the very reason that the examining psychologist judged that the testee might be more successful with performance test problems

Arthur, on the other hand reported that for 435 clinic cases, having Stanford Binet IQ's of less than 95 there was no group trend in the direction of higher IQ ratings on the Arthur Performance Scale, even among the duller individuals in the group [20] She also found that for 60 mentally deficient cases (ages 15 to 20) who were routinely examined the differences between verbal and performance IQ s were, on the whole negligible These are significant findings with regard to

[20] Grace Arthur "An Attempt to Sort Children with Specific Reading Disability from Other Non Readers *Journal of Appl ed Psychology* Vol 11 1927 pp 251 264 "The Relative D fficulty of Various Tests for Sixty Feeble minded Individuals *Journal of Cl nical Psychology* Vol 6 1950 pp 276–279

group trends and group generalizations But they do not alter the fact that in *some individual cases*, however few, a performance scale might yield sufficiently discrepant results to preclude a diagnosis without further study of the case A discrepancy between an individual s rating on a performance scale and that on a verbal scale should be useful, rather than otherwise, in a clinical situation because then the psychologist must find the reasons for and seek an interpretation of the discrepancy This necessity prevents the formulation of unwarranted conclusions, and it results in a fuller understanding of the person being examined

Among the advantages reported for performance tests are these (1) since they do not require the use of language, individuals do not "block" as a result of feelings of inadequacy due to lack of formal schooling, (2) since all elements of the problem are visually present, some individuals proceed with greater confidence

Clinical psychologists are agreed that, where indicated, the use of performance scales can provide more information than just a rating in the form of a numerical index These tests provide an opportunity to observe qualitative aspects of behavior under standardized conditions in a variety of problem-situations A subject s approach to a problem might reveal, for example, a state of depression or agitation, hesitation or impetuousness, thoughtful deliberateness, bull headed persistence, or easy discouragement, an insightful approach or one of haphazard trial-and-error

Performance tests also have their disadvantages and limitations As already stated, they are limited in range of mental functioning tested, hence they do not differentiate well among individuals of better than average levels In fact, most performance scales thus far developed do not differentiate adequately among a large portion of the members of a representative population above twelve or thirteen years of age On the whole, it is at the lower age levels and the lower mental levels that performance tests are most useful, in addition to their usefulness with persons having the handicaps already mentioned

Geometric form boards picture completion boards, object assemblies, etc , are within almost universal experience of American school children, and since they make demands upon the significant mental processes already indicated, they may be regarded as, in some degree, measures of intelligence at these earlier age levels However, as performance tests go higher in age levels and difficulty levels, they present

problem situations which are highly specialized and for the handling of which the subjects have by no means had a common background This is especially true of the more complex and subtle form boards (e g Carl Hollow Square), performance on which is facilitated by training and experience in tasks requiring spatial perception, as in some types of engineering cabinet making and the like Since these tests do not require, to a significant degree the use of ability to make abstractions and to deal with concepts they fail to measure some of the most important aspects of mental activity

The reader will have noted that not all authors of performance tests agree as to what their characteristics should be The Arthur scale was constructed to provide a nonverbal substitute for the Binet revisions Hence it was expected to have a very significant correlation with these revisions and standardization proceeded on that principle The Cornell Coxe scale is intended not as a substitute for Binet revisions but as a supplement to them Thus this performance scale was constructed on the principle that there should be a relatively low correlation between it and the verbal type of intelligence tests Experimental evidence suggests that at least in their present stage of development, performance scales are most properly and advantageously used in the manner advocated by Cornell and Coxe For performance scales are instruments with which we may test development of insightful behavior involving visual perception rather than through the use of symbols (language and number) which are essential for abstractions concept formation ideational reasoning and ability to deal with problems extending beyond one s immediate concrete environment

# 9.

## SCALES FOR INFANTS AND PRESCHOOL CHILDREN

IN THIS chapter we shall present several representative scales devised to evaluate mental development of individuals ranging in age from one *month* to six *years* Some of these scales, for the greatest part, *are not tests as that term is commonly understood They* are rather, norms and inventories of development and behavior, grouped at their respective average age levels derived from observation of children's behavior and from experimentation in a variety of situations All are administered individually, of course

### GESELL DEVELOPMENTAL SCHEDULES

These are a product of systematic study of infants and young children at the Yale Clinic of Child Development The first schedule,[1] an early effort published in 1925, provided rather crude norms at the following age levels 4, 6, 9, 12, 18, 24, 36, 48, and 60 months

At each level the inventory of activities was divided into four categories of behavior (1) *motor*, (2) *adaptive* (3) *language*, (4) *personal social* Although the normative schedules themselves have undergone considerable revision and refinement since their first appearance, these four categories, with some minor variations in terminology and analysis at times, have remained throughout Motor behavior is said to be of value * because it has so many neurological implications, and because motor capacities of the child con-

[1] A. Gesell *The Mental Growth of the Preschool Child* New York Macmillan, 1925

stitute the natural starting point for an estimate of his maturity" In adaptive behavior " we reckon with the finer sensori motor adjustments to objects and situations the coordination of eyes and hands in reaching and manipulation, the capacity to initiate new adjustments in the presence of simple problem situations which we set before the infant" Language behavior, broadly used, includes ' all visible and audible forms of communication, whether by facial expression, gesture, postural movements, vocalizations, words, phrases, or sentences [It] includes mimicry and comprehension of the communications of others" Personal social behavior " comprises the child s personal reactions to the social culture in which he lives [bladder and bowel control, feeding abilities sense of property, self dependence in play, cooperativeness, responsiveness to training and social conventions]"[2]

**The Infant Schedule.** The schedules of 1925 were followed by reports of further investigations upon which revised normative inventories of behavior were based One schedule has been devised for the examination of infants between ages of four *weeks* and fifty six *weeks* At the four week level, the inventory of behavior includes analysis of head control, arm hand posture, leg foot posture, body posture and progression regard, prehension language and social behavior At the fifty six week level the inventory includes the following categories body posture and progression, prehension, manipulation and adaptation, language and social behavior[3] Each of these categories of behavior is evaluated in the case of a particular infant, by observing him in a number of situations[4] Each situation is broken down into a number of possible activities detailing the manner in which the infant might respond (See Table 37[5]) Since the enumerated responses follow diverse trends with age, they have been designated as follows (1) *decreasing trend*, if at ascending ages there is a progressive de

[2] A Gesell and C S Amatruda *Developmental Diagnosis* revised edition New York P B Hoeber 1947 pp 5–6

[3] A Gesell and H Thompson *The Psychology of Early Growth* New York Macmillan 1938 pp 147 ff

[4] For example, activity with a ball a bell rattle cubes cup spoon form board mirror, boxes pellet and bottle dangling ring paper and crayon patterns of body posture locomotion spontaneous activities in various daily situations such as in toilet, bath and crib kinds of play; responses to people, kinds of vocalization

[5] From Gesell and Thompson *The Psychology of Early Growth* New York Macmillan 1938 p 127 (By permission.)

TABLE 37

*Dangling Ring Behavior (4 weeks–28 weeks)*
*Situation Dangling Ring (RD)*

| RD | *Behavior Items* | 4 | 6 | 8 | 12 | 16 | 20 | 24 | 28 |
|---|---|---|---|---|---|---|---|---|---|
| (1) | Regards after delay | 77 | 54 | 64 | 65 | 27 | 13 | 14 | 5 |
| (2) | Regards immediately | 26 | 46 | 36 | 35 | 68 | 97 | 96 | 95 |
| (3) | Regards momentarily | 53 | 85 | 71 | 38 | 35 | | | |
| (4) | Regards prolongedly | 47 | 43 | 29 | 62 | 87 | 47 | 38 | 5 |
| (5) | Regards consistently | | | | | 17 | 26 | 59 | 90 |
| (6) | Disregards in midplane | 77 | 39 | 46 | 46 | 14 | | | |
| (7) | Regards in midplane | 29 | 61 | 54 | 54 | 86 | | | |
| (8) | Regards in midplane (long head) | 22 | 25 | 12 | 50 | 83 | | | |
| (9) | Regards in midplane (round head) | 32 | 75 | 70 | 56 | 88 | | | |
| (10) | Regards ring in hand | | | | | 66 | 82 | 100 | 100 |
| (11) | Regards string | | | | | 7 | 13 | 46 | 53 |
| (12) | Shifts regard | 94 | 100 | 100 | 96 | 93 | 46 | 38 | 41 |
| (13) | Shifts regard to surroundings | 75 | 68 | 61 | 35 | 13 | 16 | 14 | 5 |
| (14) | Shifts regard to Examiner's hand | 28 | 64 | 61 | 77 | 48 | | | |
| (15) | Shifts regard to Examiner | 41 | 54 | 57 | 65 | 64 | 27 | 24 | 27 |
| (16) | Shifts regard to hand | 0 | 4 | 7 | 8 | 19 | 5 | 3 | |
| (17) | Follows past midplane | 44 | 62 | 50 | 58 | 84 | | | |
| (18) | Follows past midplane (lg h) | 20 | 33 | 25 | 37 | 83 | | | |
| (19) | Follows past midplane (rd h) | 55 | 75 | 60 | 67 | 77 | | | |
| (20) | Follows approximately 180° | 16 | 43 | 46 | 50 | 68 | | | |
| (21) | Follows approximately 180° (lg h) | 0 | 11 | 25 | 25 | 83 | | | |
| (22) | Follows approximately 180° (rd h) | 36 | 55 | 55 | 61 | 62 | | | |
| (23) | Approaches | 0 | 0 | 11 | 12 | 62 | 89 | 96 | 100 |
| (24) | Approaches after delay | | | | | 58 | 30 | 19 | 9 |
| (25) | Approaches promptly | | | | | 32 | 66 | 81 | 91 |
| (26) | Arms increase activity | 0 | 4 | 11 | 42 | 64 | | | |
| (27) | *Arms separate* | 0 | 0 | 4 | *15* | *17* | *19* | 7 | |
| (28) | Approaches with one hand | 0 | 0 | 4 | 12 | 20 | 24 | 39 | 55 |
| (29) | Approaches with both hands | 0 | 0 | 0 | 0 | 50 | 76 | 82 | 77 |
| (30) | Approaches with arms flexed | 0 | 0 | 0 | 12 | 44 | 60 | 54 | 14 |
| (31) | Hands come together | 0 | 0 | 0 | 8 | 20 | 38 | 11 | 5 |
| (32) | Contacts ring | 3 | 4 | 4 | 15 | 43 | 81 | 100 | 100 |
| (33) | Dislodges ring on contact | 3 | 4 | 4 | 8 | 20 | 35 | 28 | 5 |
| (34) | Grasps | 0 | 0 | 0 | 8 | 22 | 73 | 96 | 100 |
| (35) | Grasps after delay if grasps | | | | | | 75 | 46 | 14 |
| (36) | Grasps interdigitally | | | | | | 61 | 45 | 7 |
| (37) | Retains entire period | | | | | 20 | 19 | 40 | 65 |
| (38) | Holds with both hands | | | | | 10 | 33 | 56 | 67 |
| (39) | Hand opens and closes on ring | | | | | 30 | 11 | 10 | 14 |
| (40) | Brings ring to mouth | | | | | 38 | 58 | 82 | 74 |

TABLE 37 (continued)

| RD | *Behavior Items* | 4 | 6 | 8 | 12 | 16 | 20 | 24 | 28 |
|---|---|---|---|---|---|---|---|---|---|
| (41) | Free hand to midplane | | | | | 25 | 51 | 56 | 84 |
| (42) | Transfers | | | | | 3 | 18 | 41 | 74 |
| (43) | Drops | | | | | 78 | 56 | 41 | 32 |
| (44) | Drops immediately | | | | | 42 | 32 | 7 | 0 |
| (45) | Regards dropped ring if drops | | | | | 10 | 37 | 43 | 100 |
| (46) | (If drops) pursues dropped ring | | | | | 7 | 16 | 29 | 100 |
| (47) | (If drops) resecures dropped ring | | | | | 7 | 5 | 29 | 60 |
| (48) | Rolls to side | 3 | 4 | 8 | 4 | 35 | 42 | 38 | 18 |
| (49) | Frets | 9 | 14 | 4 | 8 | 27 | 23 | 32 | 21 |

crease in percentage of infants showing that behavior, (2) *increasing trend*, if at ascending ages there is a progressive increase in percentage showing that behavior, (3) *focal trend*, if at consecutive ages there is an increase, followed by a decrease in percentage giving that response The "increasing" and "decreasing" behavior items were allocated to age levels on the basis of fifty percent frequency The "focal" behavior items were placed at age levels at which they are most frequently observed [6]

**Scoring.** The infant's responses are scored plus or minus, depending upon whether or not he manifests the enumerated behaviors The score on each item of behavior is then noted on a record sheet in accordance with the categories listed above The infant's "distinctive" (modal) level of behavior is found by observation, from this level, responses showing greater or lesser degrees of maturity are counted, an algebraic sum of the deviating responses is found, this sum is then related to the "distinctive" level in order to determine whether the trend is in a plus or minus direction Finally, a rating is assigned the infant in each of the categories, thus providing a profile of development The following scores are illustrative of the ratings that might be found in a particular case [7]

| | | |
|---|---|---|
| postural behavior | 28− | weeks |
| prehensory behavior | 28+ | weeks |
| perceptual behavior | 28 | weeks |
| adaptive behavior | 28 | weeks |
| language behavior | 32 | weeks |

[6] A fourth type of item was also found those having a *fluctuating trend* that is having more than one focus But they are disregarded in the scoring

[7] For a detailed account of scoring method see Gesell and Thompson, *op cit* pp 209 ff

### *15 Month Level*

*Motor*

Walks few steps, starts and stops
Walks falls by collapse
Walks has discarded creeping
Stairs creeps up full flight
Cubes tower of 2
Pellet placed in bottle
Book helps turn pages

*Adaptive*

Cubes tower of 2
Cup and Cubes 6 in and out cup
Drawing *incipient imitation stroke*
Formboard places round block
Formboard adapts round block promptly

*Language*

Vocabulary 4–6 words or names
Jargon uses
Book pats picture
Picture card points to dog or own shoe

*Personal Social*

Feeding has discarded bottle
Feeding inhibits grasp of dish on tray
Toilet partial toilet regulation
Toilet bowel control
Toilet indicates wet pants
Communication says 'ta ta' or equivalent
Communication indicates wants (points or vocalizes)
Play shows or offers toy to mother or examiner
Play casts objects playfully or in refusal

### *72 Month Level*

*Motor*

Jumps from height of 12 , landing on toes only
Advanced throwing
Stands on each foot alternately, eyes closed
Walks length of 4 cm board
Copies diamond

*Adaptive*

Builds 3 steps with cubes
Draws man with neck, hands on arms, and clothes

Draws man with 2 dimensional legs
Copies diamond
Adds 9 parts to incomplete man
Discriminates 5 weights no error
Detects missing parts of pictures
Repeats four digits
Gives correct number of fingers on single hand and on both
Adds and subtracts within five

*Language*

Binet items used here

*Personal Social*

Ties shoe laces
Differentiates A M and P M
*Knows right and left or complete reversal*
Recites numbers up to the thirties

**Scoring** These schedules are not scored quantitatively They are a general clinical guide intended for use in estimating the developmental status of a given child in respect to the four designated categories of behavior

**Validity and Reliability** In respect to validity and reliability Gesell and his collaborators take the same position presumably as that quoted in connection with their other schedule for no statistical evidence is provided beyond percentages passing at the several age levels

Application of the schedules is a simple matter of determining how well a child s behavior fits one age level constellation rather than another by the method of direct comparison There is nothing mathematical in this determination neither is there anything mystical about it It amounts to matching which is neither calculation nor intuition [11] Performance and developmental status are reported separately for each of the four categories of behavior in terms of the four approximate age levels

**Evaluation** Inspection of these two developmental schedules reveals that each is a combination of some aspects of mental development

[11] *Ibid* p 320 While we believe there should be nothing mystical about validity and reliability of a psychological scale we also believe that subjective comparison and intuition should not be substituted for scientific determination of reliability and validity

(as usually understood), motor development, sensory development and perception, and development of personal habits (often called social development)

The schedule for infants (4 weeks to 56 weeks of age) has value for the experienced psychologist because it provides means, experimentally and clinically derived, of estimating, nonquantitatively, specified aspects of a child s development within the first year of life But as psychological tests, this schedule does not satisfy the demands of standardization in terms of norms, reliability, and validity The population sample was small (49 boys and 58 girls) and restricted (from a homogeneous middle class background) It yet remains for some psychologists to subject the schedule to rigorous studies of reliability and validity before it can be used with considerable confidence for predictive purposes On the positive side, it can be said that superior experimental techniques were used, and a great deal of careful observation, experience, and behavioral insight went into the derivation of the developmental schedule For these reasons, when applied by skilled observers, it is useful in appraising an infant's developmental status as it appears *at the time of the examination*

For the reasons stated above, the second schedule of development and behavior (for ages 15 to 72 months) is of questionable value For this group of children, especially those two years of age and older, there are other scales that have been standardized and that can be used with more confidence With some of these, the reader is already familiar (e g , the Stanford Binet), others will be described in the following pages

## MINNESOTA PRESCHOOL SCALE[12]

This scale, in two forms, is an adaptation and restandardization of test items chosen from the earlier work of a number of psychologists, plus some original additions It is designed for use with children from age eighteen months to six years

The scale includes the following twenty-six tests pointing out parts of the body, pointing out parts in pictures, naming familiar objects, copying a circle, triangle, and diamond, imitative drawing (vertical

[12] Developed by F L. Goodenough J C Foster and M J Van Wagenen Published by Educational Test Bureau, 1932 and 1940 The revised test manual 1940, is by F L. Goodenough, K M Maurer, and M J Van Wagenen

and horizontal strokes and a vertical cross), block building response to pictures Knox cube imitation (tapping a series of cubes in a given order), obeying simple commands, comprehension ("What should you do when you are hungry?"), discrimination of geometric forms naming objects from memory, recognition of forms color naming tracing a form, picture puzzles (object assembly), incomplete pictures, digit span picture puzzles diagonal series (more difficult object assembly), paper folding, absurdities (verbal), mutilated pictures vocabulary, word opposites imitating position of clock hands speech (length of sentence spoken by child during examination)

*Scoring.* The raw score is converted into what is known as a C-score which in turn can be converted into an IQ equivalent by means of tables provided in the manual [13] The Minnesota scale also provides for another type of score known as "percent placement," which is defined as the percentage of the difference between the score of the most backward and the score of the most advanced child likely to be found in a representative group of a thousand children of similar age Thus if the lowest C-score in a given group is 50 and the highest is 110 the range is 60 points A child who gets a C-score of 65 is 15 units or 25 percent (15/60), above the lowest score in the group His "percent placement" score, then is 25

The norms of this scale are so arranged that it is possible to obtain three separate scores for children above thirty months of age a verbal a nonverbal, and a total score For a child under thirty months of age only the total score is used because the authors of the scale were unable to work out a system of differentiated scoring for these earlier levels A rough analysis is possible, however, to determine whether a pronounced difference between verbal and nonverbal responses exists If such a difference is found at any age within the range of the scale then, as the case may be, handicap or acceleration in respect to language or perceptual motor ability may be inferred

[13] The C-score " represents the difficulty of the tasks with which [a child] may be expected to succeed in 50 percent of his trials" (*Manual* p 91) It is a form of "absolute scaling," the units of which presumably increase by steps that are approximately equal in difficulty It is a variation on the familiar standard-score technique For a more detailed description of the C-scores and the basis of determining IQ equivalents see F L. Goodenough and K. M. Maurer *The Mental Growth of Children From Two to Fourteen Years* Minneapolis University of Minnesota Press, 1942, Chapter IV

**Validity and Reliability.** The Manual of the Minnesota scale does not provide data specifically designated as evidence of validity We may infer, however, that the authors regarded the following facts as their basis of validity (1) the adaptation and use of types of test items considered by many psychologists, over a period of years, to have validity, (2) a standardization group of 900 children, ranging in age from eighteen months to six years (100 in each of nine half-year age groups), who were balanced equally as to sex and whose

TABLE 38

*Correlations Between Minnesota Preschool and Stanford Binet IQ s*

| *Age in Months at Taking Minnesota* * | *Correlations* 1916 S B | 1937 S B |
|---|---|---|
| Under 36 | 45 | 21 |
| 36–47 | 64 | 61 |
| 48 and over | 65 | 68 |

* The number of cases in each group was large ranging from 141 to 841 ) From Goodenough and Maurer op cit Part II

fathers were representative of the distribution of occupational levels in the general population

In another and more recent publication, data relevant to the scale s validity are available [14] Children who had been tested originally with the Minnesota scale at various ages during their preschool years were retested with the 1916 Stanford Binet in some instances or with the 1937 revision in others The intervals between tests and retests varied from a few months to about ten years

When the 1916 revision was used in retesting and the results were correlated with the total scores of the Minnesota, the range of coefficients for the various groups was from a low of 25 to a high of 75

When the 1937 Stanford Binet was used in retesting the correlations with original Minnesota scale total scores yielded coefficients from 15 to 76

Table 38 shows the median correlations between original Minnesota IQ equivalents found at various ages, and the retest Stanford-Binet IQ s at ages ranging from 4½ to 13½ years

[14] Goodenough and Maurer *op cit* Part II

If we accept the Stanford-Binet scales as significant criteria of validity, as many psychologists have in actual practice, then we must conclude that the Minnesota scale has low validity below the age of thirty-six months, but that it has much greater validity for individuals above three years of age

In this connection two considerations must be kept in mind First, it has been found that all scales devised for use with children below the age of eighteen months show a low or very moderate correlation with retest results in later childhood The probable reasons for this fact will be presented later Second, the correlation coefficients between results of testing and retesting the same subjects with the *same* scale tend to decrease somewhat as the time interval between examinations increases

Reliability data of the Minnesota scale are variable at the different age levels The coefficients between the C-scores on the two forms of the scale (that is, the test retest method), with intervals of one to seven days, were 68 to 94 for the verbal tests, 67 to 92 for the nonverbal tests, and 80 to 94 for the combined total scores The average reliability coefficients for a single form, within an age range of six months, were 86 for the verbal, 82 for the nonverbal, and 89 for the total scores

## CATTELL DEVELOPMENTAL AND INTELLIGENCE SCALE[13]

This scale, of superior merit, covers the range from two to thirty months Its test items are adaptations of many which were developed and included in earlier tests, notably those of Gesell and his associates Cattell states the scale " has been so constructed as to constitute an extension downward of Form L of the Stanford Binet tests Between the ages of twenty two and thirty months Stanford Binet items are intermingled with other items Thus, using the infant test items for the early months and the Stanford Binet tests for the older ages with a mixture of the two between, one continuous scale from early infancy to maturity has been attained "[14] The test items are grouped at age levels as they are in the Stanford Binet. Groupings are provided at each month from two through twelve, at two-month intervals in the second year, and at twenty-seven and thirty months

[13] Psyche Cattell *The Measurement of Intelligence of Infants and Young Children* New York Psychological Corporation 1940

[14] *Ibid* p 24

The following three age levels illustrate the nature of the items and their arrangement

*Two months*

(1) Attends voice
(2) Inspects environment
(3) Follows ring in horizontal motion
(4) Follows moving person
(5) Babbles
(Alt a) Follows ring in vertical motion
(Alt b) Lifts head in prone position

*Ten months*

(1) Uncovers toy
(2) Combines cup and cube
(3) Attempts to take third cube
(4) Hits cup with spoon
(5) Pokes fingers in holes of peg board
(Alt ) Picks up spoon before cup

*Thirty months*

(1) Differentiates bridge from tower
(2) Imitates drawing lines and circles
(3) Stanford Binet three hole form board rotated
(4) Folds paper
(5) Stanford Binet identifying objects by use
(Alt a) Identifies p ctures from name
(Alt b) Concept of one

The scale was standardized by longitudinal testing 1346 examina tions were made on 274 children at the ages of three six nine, twelve, eighteen twenty four thirty and thirty six months [1]

In the process of standardization it was Cattell s purpose among other things to improve on earlier scales by (1) improving objective procedures for administering and scoring (2) eliminating items of the personal social category which are markedly influenced by home training (3) eliminating items which are indicators of large motor control (4) providing more accurate age scaling (5) provid

[1] It should be noted that wh le the scale was standard zed only on these age levels groups of items are provided at certain age levels between them The placement of items between the standardizat on levels was estimated Cattell states however that the ind cations are the scale may be used with only a little less accuracy with children between the standardization ages At the same time she urges the exercise of caution in interpreting test results at the ages between the standardization levels

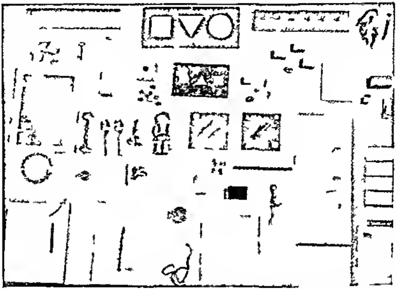


FIG 9 1 *Complete set of material for administering the infant tests From P Cattell The Measurement of Intelligence of Infants and Young Children Psychological Corporation (By permission )*

ing an adequate age range so that continuity of development can be studied (6) providing a more nearly equal distribution of items over the age range covered

FIG 9 2 *Regards Cube Age 3 Months*

Material A one inch cube painted bright red

Procedure *As the child is sitting in an upright position before the table the cube is placed on the table within easy view of him The cube may be tapped on the table or moved about to attract the child s attention*

Scoring *Credit is given if the child observes the cube His eyes must remain on or return to the cube after the examiner has removed his hand In other words the examiner must make sure that it is the cube and not his hand which is observed*

From P Cattell *The Measurement of Intelligence of Infants and Young Children* Psychological Corporation (By permission )

FIG 9 3 *Picks Up Spoon Age 5 Months*

Material *Teaspoon*

Procedure *The spoon is placed directly in front of the child (sitting position) within easy reach*

Scoring *Credit is given if the child makes a definite effort to reach for and pick up the spoon and succeeds but if the spoon is picked up by reflex closure of the hand on chance contact it is not credited Accurate reaching however is not to be expected at this age*

From P Cattell *The Measurement of Intelligence of Infants and Young Children Psychological Corporation (By permission )*

Scoring The method of scoring is the same as that used with the Stanford Binet scale Each item is scored as either plus or minus no partial credits are given Since there are five items at each age level the credit given for each item passed is one fifth of the interval cov-

FIG 9 4 *Places Round Block in Form Board Age 16 Months*

Material *The form board is similar to Gesell s It is made of a three-eighths inch board 36 × 16 cm stained dark green Three holes are cut in the board equidistant from each other and from the edges From left to right the holes are a circle 8 7 cm in diameter an equilateral triangle with sides 9 3 cm and a square with sides 7 5 cm The inserts are made of wood 2 cm thick and painted white The circle is 8 5 cm in diameter the sides of the triangle 9 cm and those of the square 7 3 cm*

Procedure *The form board is placed before the child with the circle on his left and the base of the triangle toward him The circle is placed in its recess and the child is allowed to take it out then he is asked (with appropriate gestures) to Now put it back*

Scoring *Credit is given if the child replaces the round block If it is done with an evidently purposeful act one trial is enough but if there is some doubt as to whether or not it was a chance replacement no credit should be given unless it is placed a second time (Credit is given for replacing the round block in the reversed board at eighteen months )*

From P Cattell *The Measurement of Intelligence of Infants and Young Children Psychological Corporation (By permission )*

ered by the particular series of tests Thus, in a series of items spanning a one month interval, each item passed carries credit of 2 of a month, when the interval is two months, the credit per item is 4 of a month, with a three month interval, it is 6 of a month The Cattell scale, like the Stanford-Binet, uses a basal age and sums up the credits at higher levels to obtain a mental age, and from that an IQ

**Validity and Reliability.** Although percent passing each item at successive ages was used as evidence of validity, the principal criterion was the correlation between Cattell scale IQ ratings, obtained up to the age of thirty months, and Stanford-Binet (1937) IQ's obtained

TABLE 39

*Validity Coefficients Cattell and Stanford Binet Scales*[18]

| *No* | *Ages at Examinations* | *Coefficients* |
|---|---|---|
| 42 | 3 mos and 36 mos | 10 ± 10 |
| 49 | 6 " " " " | 34 ± 08 |
| 44 | 9 " " " " | 18 ± 10 |
| 57 | 12 " " " " | 56 ± 06 |
| 52 | 18 " " " " | 67 ± 05 |
| 52 | 24 " " " " | 71 ± 05 |
| 42 | 30 " " " " | 83 ± 03 |

with the same children at the age of thirty-six months Table 39 shows the validity coefficients

It is obvious that, accepting the Stanford-Binet as the criterion, the coefficients are practically negligible for tests given during the first nine months of life In this respect they are much the same as other current scales For the later ages, up to thirty months, the coefficients increase very appreciably and are on the whole superior to those found with most other scales designed for these age levels

In spite of the low predictive value of the coefficients at the earlier ages, Cattell has found, from study of individual cases, that the tests may be of considerable assistance to the clinician in appraising infants who are marked deviants from the norm This is the case especially with infants who get a high quotient, for, Cattell reports, they have

[18] From Psyche Cattell *The Measurement of Intelligence of Infants and Young Children* New York Psychological Corporation 1940, p 49 (By permission.)

appreciably better than average chances of earning a high rating at the age of two or three years

Reliability of the scale was calculated by the odd-even procedure and corrected by the Spearman Brown formula Coefficients ranged from a low of 56 ± 05 at the age of three months to a high of 90 ± 01 at eighteen months The median coefficient was 86 ± 02 These coefficients compare favorably with those found for other scales

## MERRILL-PALMER SCALE OF MENTAL TESTS[19]

Although the norms of this scale are based upon 631 cases ranging in age from eighteen to seventy seven months its author does not recommend its use with children below twenty four months or above sixty three months of age

The scale consists of ninety three items arranged in order of diffi culty There is no attempt to group the tests according to types of function or behavior involved The age norm (called age at par ) for each item is given this being the age at which fifty percent of the children were successful Although there are ninety three items there are only thirty-eight *different* items Some (twenty one) recur several times at different age levels at later ages a higher level of perform ance is required (in terms of quality or quantity of response or in rate of activity) if credit is to be earned These are called variable score tests by the author Other items (seventeen) occur only once in the scale they are called all or none tests

The scale tests some language (e g What runs? What scratches? These are known as action agent tests Also simple ques tions like What does a doggie say? ) manipulat on of the body (e g opposition of thumb and fingers crossing feet) motor skills and coordination (e g throwing a ball buttoning) visual insights (e g building with blocks copying a circle and a cross completing form boards and picture puzzles) and recognizing familiar objects and colors

Scoring A point or raw score is obtained first This may be con verted into one or more of several relative indexes namely (1) men

---

[19] Rachel Stutsman *Mental Measurem nt of Presch ool Ch ldren* Yonk rs N Y World Book 1931

tal age, (2) standard deviation on the basis of point scores, (3) staodard deviation on the basis of mental ages, (4) standard deviation oo the basis of IQ s, (5) and percentile raoks on the basis of raw scores It is suggested, however, that use of the IQ itself is in advisable in connection with the Merrill Palmer scale because its IQ deviations are not the same or close eoough io size at the various age levels

Stutsman provides and suggests the use of a 'Guide for Personality Observatioos' in connection with this scale While these observations do oot affect the scoring they are, nevertheless, useful to the clinician in interpreting a child s responses during the examinatioo The following traits are observed and rated, as they are manifested during testing self reliance, self-criticism, irritability toward failure, degree of praise needed for effective work, initiative and independence of action self-consciousness spontaneity and repression, imaginative tendencies, reaction type (slow and deliberate, calm and alert, quick and impetuous), speech development dependence on parent, and other observations Value of these observations and ratings, obviously will be dependent on the skill and experience of the examiner These and similar observations, as already pointed out, are desirable, in fact essential in the complete report on and evaluation of any individual s test performance

**Validity and Reliability** Criteria of validity were those generally used (1) known groups, (2) ratings by oursery school staff, (3) small overlapping of distribution of total scores between age groups (4) correlation with chronological age ($r = 92 \pm 004$), and (5) correlation with the Stanford Binet ($r = 79 \pm 019$, 159 children in the staodardization group, between 3 aod 6 years of age) The correlation coefficient for the last criterion must be interpreted in the light of the fact that the age range was three years

In the guide describing the Merrill Palmer scale and its standardization, no data on its reliability are provided Subsequeot studies, however, furnish information from which we may infer its reliability When Stutsman retested a group of 77 children (ages 2 to 5 years) after ao ioterval of two months, she found a correlation coefficient of 72 between the two sets of scores Wellman, retesting a group of 44 childreo (ages 20–62 months) after an interval of one week,

found a correlation coefficient of 92 between the scores of the first and second tests [20]

## EVALUATION OF SCALES FOR INFANTS AND PRESCHOOL CHILDREN

**Technical Problems** Time or speed scores should not be used in tests for these age levels The measurement of rate of performance is inadvisable and can be misleading for at least two reasons (1) speed of performance has not yet become a motivating factor in very young children, (2) the shifting attention of children at these age levels can obscure their true levels of skill and insight

The grouping of test items according to types of activity, as in the Gesell schedule has the advantage of readily indicating functions that are retarded and those that are accelerated in the case of the child being examined While this kind of analysis is not so immediately apparent in an age scale, like Cattell s it is nevertheless possible

Since most of the usual validity criteria are not available in standardizing tests for infants and young children, this technical problem can be solved only through longitudinal studies, following up the same individuals over a considerable span of years, and correlating early test performance with later acceptable criteria of validity Some efforts have been made in this direction [21]

Determination of item and subtest reliability, within a scale presents difficulties too If the odd-even method is used the results can be affected by the fluctuating attention of the subject If the test retest method is used, the results can be affected by the irregularity of

[20] B L Wellman *The Intelligence of Preschool Children as Measured by the Merrill Palmer Scale of Performance Tests* University of Iowa Studies in Child Welfare Vol 15 No 3 1938

[21] See V L Nelson and T W Richards Studies in Mental Development I Performance on Gesell Items at Six Months and Its Predictive Value for Performance on Mental Tests at Two and Three Years" *Journal of Genetic Psychology* Vol 52 1938 pp 303 325 II "Analysis of Abilities Tested at the Age of Six Months by the Gesell Schedule" *ibid* pp 327 331 III "Performance of Twelve Months Old Children on the Gesell Schedule and Its Predictive Value for Mental Status at Two and Three Years" *ibid* Vol 54 1939 pp 181 191 "Abilities of Infants During the First Eighteen Months" *ibid* Vol 55 1939 pp 299 318 Also N Bayley "Consistency and Variability in the Growth of Intelligence from Birth to Eighteen Years" *ibid* Vol 75 1949 pp 165–196

tal age, (2) standard deviation on the basis of point scores, (3) standard deviation on the basis of mental ages, (4) standard deviation on the basis of IQ s, (5) and percentile ranks on the basis of raw scores It is suggested, however, that use of the IQ itself is inadvisable in connection with the Merrill-Palmer scale because its IQ deviations are not the same or close enough in size at the various age levels

Stutsman provides and suggests the use of a "Guide for Personality Observations' in connection with this scale While these observations do not affect the scoring, they are, nevertheless, useful to the clinician in interpreting a child s responses during the examination The following traits are observed and rated, as they are manifested during testing self-reliance, self-criticism, irritability toward failure, degree of praise needed for effective work, initiative and independence of action, self-consciousness, spontaneity and repression, imaginative tendencies, reaction type (slow and deliberate, calm and alert, quick and impetuous), speech development, dependence on parent, and other observations Value of these observations and ratings, obviously, will be dependent on the skill and experience of the examiner These and similar observations, as already pointed out, are desirable, in fact essential, in the complete report on and evaluation of any individual s test performance

**Validity and Reliability.** Criteria of validity were those generally used (1) known groups, (2) ratings by nursery school staff, (3) small overlapping of distribution of total scores between age groups, (4) correlation with chronological age ($r = 92 \pm 004$), and (5) correlation with the Stanford Binet ($r = 79 \pm 019$, 159 children in the standardization group, between 3 and 6 years of age) The correlation coefficient for the last criterion must be interpreted in the light of the fact that the age range was three years

In the guide describing the Merrill Palmer scale and its standardization, no data on its reliability are provided Subsequent studies, however, furnish information from which we may infer its reliability When Stutsman retested a group of 77 children (ages 2 to 5 years) after an interval of two months, she found a correlation coefficient of 72 between the two sets of scores Wellman, retesting a group of 44 children (ages 20–62 months) after an interval of one week,

found a correlation coefficient of 92 between the scores of the first and second tests [20]

## EVALUATION OF SCALES FOR INFANTS AND PRESCHOOL CHILDREN

**Technical Problems.** Time or speed scores should not be used in tests for these age levels The measurement of rate of performance is inadvisable and can be misleading for at least two reasons (1) speed of performance has not yet become a motivating factor in very young children, (2) the shifting attention of children at these age levels can obscure their true levels of skill and insight

The grouping of test items according to types of activity, as in the Gesell schedule has the advantage of readily indicating functions that are retarded and those that are accelerated in the case of the child being examined While this kind of analysis is not so immediately apparent in an age scale like Cattell s, it is nevertheless possible

Since most of the usual validity criteria are not available in standardizing tests for infants and young children this technical problem can be solved only through longitudinal studies following up the same individuals over a considerable span of years, and correlating early test performance with later acceptable criteria of validity Some efforts have been made in this direction [21]

Determination of item and subtest reliability within a scale, presents difficulties too If the odd-even method is used the results can be affected by the fluctuating attention of the subject If the test retest method is used the results can be affected by the irregularity of

---

[20] B L Wellman *The Intelligence of Preschool Children as Measured by the Merrill Palmer Scale of Performance Tests* University of Iowa Studies in Child Welfare Vol 15 No 3 1938

[21] See V L Nelson and T W Richards Studies in Mental Development I Performance on Gesell Items at Six Months and Its Predictive Value for Performance on Mental Tests at Two and Three Years *Journal of Genetic Psychology* Vol 52 1938 pp 303 325 II Analysis of Abilities Tested at the Age of Six Months by the Gesell Schedule *ibid* pp 327–331 III Performance of Twelve Months Old Children on the Gesell Schedule and Its Predictive Value for Mental Status at Two and Three Years *ibid* Vol 54 1939 pp 181–191 Abilities of Infants During the First Eighteen Months" *ibid* Vol 55 1939 pp 299 318 Also, N Bayley "Consistency and Variability in the Growth of Intelligence from Birth to Eighteen Years" *ibid* Vol 75 1949 pp 165–196

growth tempos, when the time interval is significant Significance of the time interval varies with the age of the subject the younger the child, the shorter the significant interval The desirable procedure would be to make retests within a week

Although most scales for these early age levels extend upward beyond the two and three year levels, many psychologists recommend the use of the Stanford Binet from age two, because of its more adequate standardization Since the Cattell scale is an extension downward of the Stanford Binet, and since it overlaps with the latter, it is a sound alternate for the Stanford Binet The Merrill-Palmer has also been found to be quite useful to the age of three or three and one-half years

**Uses.** Psychological tests for children at these ages are used for two main purposes (1) to determine a child s developmental status, with respect to the functions being evaluated, at the time of examination, and (2) to predict, so far as possible, future developmental and intellectual status Most psychologists agree that the first purpose is reasonably well satisfied With regard to the second of these purposes, the infant scales with one possible exception, used with subjects below the age of fifteen or eighteen months have not proved adequate, for when test ratings obtained in about the first year and a half of life have been correlated with subsequent ratings obtained with the Stanford Binet and other scales, the correlations found were so low as to be negligible In fact, even an occasional small negative coefficient has been found Therefore, when a child is examined within the first eighteen months of life, for purposes of predicting his future mental development and status, little weight can be placed upon the numerical rating, except in the cases of infants who deviate markedly from the average, in either direction

Cattell s scale is superior to most others in this respect. The first three coefficients ( 10 to 34) in the table showing validity coefficients of this scale are characteristic of those generally found for the first year of life But the remaining coefficients of validity are higher than those found for other preschool tests In general, predictive value of scales for preschool children increases after the age of eighteen months or two years While the many correlational studies published on preschool groups aged eighteen months or more show coefficients varying over a considerable range, many of them fall in the 40 s,

50's and 60's, with relatively few others higher or lower [22] In general, the higher the preschool age at the initial test, the higher will be the relationship of initial score to scores on retests

In one instance, at least, psychologists have been concentrating upon devising scales to test very young infants exclusively—between the ages of 4 and 36 *weeks*—when adoptions are made most frequently These scales, known as The Northwestern Infant Intelligence Tests, emphasize the child's adaptation to the physical and social environment Their ultimate value, too, will depend upon their demonstrated predictive efficiency [23]

In spite of their low predictive value for very young infants available scales are of assistance to an experienced clinical psychologist in appraising a child's behavioral and mental development when attention is given to analysis of performance on the various parts rather than to numerical scores alone, and when the analysis is used in conjunction with other clinical data Developmental and intelligence tests for preschool children must be used with more than ordinary precaution, for the value of the findings is exceptionally dependent upon the skill of the examiner in eliciting the child's best efforts and in being able to appraise his general behavior during the examination session

There are several reasons why results of tests in infancy and the earlier preschool periods do not have more value in predicting future mental status First, resistance to examiner, shyness, failure to exercise maximum effort, and other emotional conditions are undoubtedly operative in some instances More important and fundamental, however, is the fact that there are changes and irregularities in the tempo of development of numerous young children It has been found that successive examinations of individual infants show fluctuations between two or three levels, or they may show a consistent trend downward or upward before leveling off to variations within a relatively narrow range of ratings [24] These fluctuations and trends in rate of development may be due to changes in mental organization, that is,

---

[22] For a compact summary of these studies see K M Maurer *Intellectual Status at Maturity as a Criterion for Selecting Items in Preschool Tests* Minneapolis University of Minnesota Press 1946 Chapter 2

[23] By A R Gilliland and A M Shotwell Northwestern University Evanston Ill

[24] See Cattell *op cit* 52 ff also N Bayley "Mental Growth in Young Children" *Thirty ninth Yearbook* National Society for the Study of Education Bloomington Ill Public School Publishing Co 1940 Part II pp 11–47

differences in the age of appearance of various functions and differences in their rates of development—appearance of new functions and changes in rates being especially rapid in the first two years of life

Closely allied to the foregoing is the fact that tests included in infant scales are dissimilar to those used at later age levels, so that little correlation is to be expected The reader will have observed that tests used in appraising an infant's development in the first eighteen months of life are largely of relatively simple motor activities and of sensory perception These have never been found to correlate significantly with tests used at later age levels, which increasingly involve the higher and more complex mental functions (See the chapter on definitions and nature of intelligence ) It may be that psychologists will not be able to devise infancy tests having greater predictive value, for it seems that those functions which are subsumed under the term "intelligence' do not reach a measurable magnitude until an age later than infancy That is, intelligence, as psychologically understood and defined, does not emerge sufficiently during the earliest phase of development

In conclusion, then, it may be said that when tests are used with children below the age of eighteen months, emphasis should be placed upon an analysis of performances and their evaluation of the child's *present* status Thereafter, the scales increase in value for the purpose of predicting later mental level Bayley,[23] after surveying her own and other researches, concluded that tests given between two and four years of age will predict eight- and nine-year intelligence test performance with moderate success ($r = 55$), while tests given at four years of age will predict eight- and nine year performance much more satisfactorily ($r = 75$) These conclusions, however, were written before the publication of Cattell s scale and its validity data It appears, therefore, that, as Cattell has shown, it is possible to devise scales of *significantly higher predictive value for* use with *preschool* children who are above eighteen months of age [24]

[23] N Bayley *op cit* pp 16 ff

[24] It would be highly desirable to have more psychologists and institutes devote more time energy and money to research on the very important problems in this area and perhaps to divert some of the time energy and money now being lavished on the study of personality and projective methods.

# 10.

## NONVERBAL GROUP SCALES OF MENTAL ABILITY

### BEGINNINGS

The original Binet scale and its several revisions are administered to one person at a time, hence they are called individual scales This is true, also of the performance scales already described Individual scales, obviously, are time-consuming and require that the examiner be highly skilled in administering them, in interpreting responses, and in evaluating the subject s behavior during the course of the examination Impelled perhaps by the prevailing urge for "efficiency and mass production, and by a desire to investigate large-scale problems, American psychologists undertook to develop tests which could be administered to a group of persons—large or small—all at one time

World War I provided the occasion for the organization of the first group test Prior to 1917, psychologists had been experimenting with test items and organization with a view to group examination Shortly after the United States entered World War 1, a psychological branch was formed in the army in order to develop and use group scales for the purpose of general classification of soldiers on the basis of mental ability, so far as the tests might measure that trait A few other devices were developed for use in the army—e g trade tests—but in the army of World War I the work and contribution of psychologists were very largely in the testing of general ability

Although the 1916 Stanford Revision and the Yerkes Point Scale

were employed to some extent, as well as an individual performance scale, the main task in the army was one of testing very large numbers of men in a short space of time Consequently, the Army Alpha scale (verbal) and the Army Beta scale (nonverbal) were organized, both being group scales These were actually the product of the contributions of individual psychologists, notably Arthur S Otis, who pooled their experience, experimental results, and resources

About 1,750,000 men were tested in the army of World War I Though the scales were by no means highly satisfactory instruments, and though the men were very often examined under unfavorable conditions, the results obtained were of some assistance in the selection of men for advanced or special training, on the one hand, and of men of such inferior ability as to be unsuited for military training, on the other

The use of psychological testing in the army of World War I had many outgrowths, some of which were, no doubt, unforeseen by the psychologists themselves The data were reported and analyzed in a huge volume[1] On the basis of these data, many periodical articles and books appeared on such subjects as racial and national differences in intelligence, geographic differences in intelligence within the United States, differences between occupational groups, relationship between educational status and intelligence, and the general intellectual level of the American adult Not only were many of these data of doubtful validity but some of the interpretations and publications based upon them gave rise to serious misapprehensions in regard to the foregoing problems, which are loaded with social and educational implications Another result of psychological testing in the army was the impetus it gave to the development of group tests for civilian purposes, notably in educational work at all levels, from kindergarten through university Also, it set a precedent, for in World War II psychological testing was conducted on a vast scale in all departments of the armed forces

The types of test materials included in the army group scales and in the numerous group scales subsequently developed were not all innovations For example, tests of memory, sentence completion, free and controlled word association, arithmetic computation, vocabulary, classification of objects, and following directions had been in process

[1] Yerkes R M, editor, *Psychological Examining in the United States Army Memoirs of the National Academy of Sciences, Washington* Government Printing Office, 1921, Vol 15

of experimentation beginning within the last twenty five years of the nineteenth century, in the United States and in several European countries

## CHARACTERISTICS OF GROUP TESTS OF MENTAL ABILITY

With a few exceptions, group tests—implicitly or explicitly—are constructed on the principle that intelligence is a general capacity and that it should be measured by means of sampling a variety of mental activities Inspection of the scales shows, therefore, that they include in various combinations such items as following directions, arithmetical problems, practical judgment (in connection with "common sense" problems), word meaning, disarranged sentences, completion of number series, completion of sentences, verbal analogies, information, mazes, three-dimensional visualization and counting of cubes, symbol-digit combinations, picture absurdities, picture arrangement, geometrical construction ("paper form board"), and geometric pattern analogies Samples of each of these will be presented later when illustrative scales are described

In most group scales, the items of each type (e g, number series) are placed together in separate subtests or parts, beginning with the easiest and progressing by intervals—as nearly equal as may be achieved—to the most difficult The principle involved here is this by means of such an arrangement of items, every individual for whom the test is intended should be able to get some items correct and to proceed to a level of difficulty which represents his maximum in that particular type of mental activity

Occasionally, however, it will be found that items in a scale are arranged in "spiral omnibus" fashion that is, items of various types are presented in regular or irregular order instead of being grouped separately in subtests Thus, there may be a sequence of this kind one item each in number completion, arithmetical problem, vocabulary, information, analogies, etc, then the types of items, increasing in difficulty, will be repeated in the same or in a different order

Every group scale is standardized for a specified range of ages or school grades [2] Thus the particular types of items used and the levels

---

[2] Standardization for a grade range is in practice tantamount to standardization for an age range because the scale must be adequate for the spread of ages ordinarily located within those grades Furthermore, even when a scale is specified for certain school grades it provides tables of norms for the various age groups.

were employed to some extent, as well as an individual performance scale, *the main task in the army was one of testing very large numbers* of men in a short space of time Consequently, the Army Alpha scale (verbal) and the Army Beta scale (nonverbal) were organized, both being group scales These were actually the product of the contributions of individual psychologists, notably Arthur S Otis, who pooled their experience, experimental results, and resources

About 1,750,000 men were tested in the army of World War I Though the scales were by no means highly satisfactory instruments, and though the men were very often examined under unfavorable conditions, the results obtained were of some assistance in the selection of men for advanced or special training, on the one hand, and of men of such inferior ability as to be unsuited for military training, on the other

The use of psychological testing in the army of World War I had many outgrowths, some of which were, no doubt, unforeseen by the psychologists themselves The data were reported and analyzed in a huge volume[1] On the basis of these data, many periodical articles and books appeared on such subjects as racial and national differences in intelligence, geographic differences in intelligence within the United States, differences between occupational groups, relationship between educational status and intelligence, and the general intellectual level of the American adult Not only were many of these data of doubtful validity but some of the interpretations and publications based upon them gave rise to serious misapprehensions in regard to the foregoing problems, which are loaded with social and educational implications Another result of psychological testing in the army was the impetus it gave to the development of group tests for civilian purposes, notably in educational work at all levels, from kindergarten through university Also, it set a precedent, for in World War II psychological testing was conducted on a vast scale in all departments of the armed forces

The types of test materials included in the army group scales and in the numerous group scales subsequently developed were not all innovations For example, tests of memory, sentence completion free and controlled word association, arithmetic computation, vocabulary classification of objects, and following directions had been in process

[1] Yerkes R M editor *Psychological Examining in the United States Army Memoirs of the National Academy of Sciences*, Washington Government Printing Office, 1921, Vol 15

of experimentation beginning within the last twenty five years of the nineteenth century, in the United States and in several European countries

## CHARACTERISTICS OF GROUP TESTS OF MENTAL ABILITY

With a few exceptions, group tests—implicitly or explicitly—are constructed on the principle that intelligence is a general capacity and that it should be measured by means of sampling a variety of mental activities Inspection of the scales shows, therefore, that they include in various combinations such items as following directions, arithmetical problems, practical judgment (in connection with "common-sense problems), word meaning, disarranged sentences, completion of number series, completion of sentences, verbal analogies, information mazes, three dimensional visualization and counting of cubes, symbol digit combinations, picture absurdities picture arrangement, geometrical construction ( paper form board ), and geometric pattern analogies Samples of each of these will be presented later when illustrative scales are described

In most group scales, the items of each type (e g , number series) are placed together in separate subtests or parts, beginning with the easiest and progressing by intervals—as nearly equal as may be achieved—to the most difficult *The principle involved here is this* by means of such an arrangement of items, every individual for whom the test is intended should be able to get some items correct and to proceed to a level of difficulty which represents his maximum in that particular type of mental activity

Occasionally, however, it will be found that items in a scale are arranged in spiral omnibus ' fashion that is, items of various types are presented in regular or irregular order instead of being grouped separately in subtests Thus, there may be a sequence of this kind one item each in number completion arithmetical problem, vocabulary, information, analogies, etc , then the types of items, increasing in difficulty, will be repeated in the same or in a different order

Every group scale is standardized for a specified range of ages or school grades [2] Thus the particular types of items used and the levels

---

[2] Standardization for a grade range is in practice tantamount to standardization for an age range because the scale must be adequate for the spread of ages ordinarily located within those grades Furthermore even when a scale is specified for certain school grades it provides tables of norms for the various age groups

of difficulty within a scale will depend upon the group for which it is intended For instance, a group scale designed for children from kindergarten through the second grade will be almost entirely non verbal in character, except for directions, one designed for pupils in the intermediate grades will include an increasingly larger portion of abstract and conceptual items (verbal and numerical), while tests of intelligence for high-school pupils and college freshmen are very largely, some entirely, of the verbal and numerical kind

On many group scales, an individual s score is first obtained in terms of the number of points earned, that is, a raw score From a table of norms, this score is converted into a mental age from which an intelligence quotient is calculated The manuals of some group scales provide, also, tables necessary to find an individual s percentile rank for his age or grade, or both Other group scales dispense with mental ages and intelligence quotients, and give only percentile rank equivalents for the range of point scores

Group scales are scored more rigidly and more objectively than individually administered scales, such as the Stanford Binet [3] In the former, the correct response, or responses, are supplied for each item so that they can be scored by clerks or machines In the case of the Stanford Binet and similar scales, while specimens of satisfactory and unsatisfactory responses are supplied, it is frequently necessary for the examiner himself to evaluate some responses and to decide whether or not credit should be given for them This necessary exercise of judg ment, however, does not invalidate the scales, for correlational studies have shown that there is very close agreement between experienced examiners as to the scoring of given responses

Most group scales impose time limits for each of the several sub-tests, or parts Whether this fact makes a scale a test of speed of re-sponse, solely or largely, or whether the scale measures ' power (level of difficulty the individual is capable of reaching) is a question to which answers have been provided by experiment The imposition of time limits does not necessarily make a scale a test of speed of per formance, the significance of the speed factor, in affecting the total score of a person, varies with the scale used

[3] This does not mean that the group scales are therefore better instruments. In fact, inflexible scoring is a disadvantage if the examiner s purpose is to study an individual clinically and to analyze test results qualitatively as well as quan titatively

Some group scales are entirely nonverbal in content, others are entirely verbal, while still others combine the two types of items In this chapter we shall describe several representative scales of the nonverbal variety

## PINTNER-CUNNINGHAM PRIMARY TEST[4]

This scale (having alternate forms, A and B) is intended for children in kindergarten, grade 1, and the first half of grade 2 It consists of seven subtests common observation (identifying objects commonly found in the usual environment), perception of esthetic differences, identification of associated objects (knowledge of relationships between objects, such as key and lock) discrimination of size, perception of the elements that constitute a whole picture, picture completion (noting missing parts in a picture and, from a series of choices, indicating the missing part), copying designs (using a given square of dots)

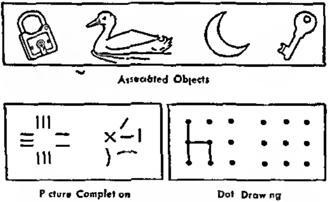



FIG 10 1 *Items from Pintner-Cunningham Primary Test* Copyright 1938 by the World Book Company (Reproduced by special permission )

Mental age norms are provided from 4 years and 1 month to 10 years

**Validity and Reliability.** As in the case of many other group scales, validity of the Pintner-Cunningham test is given in terms of its cor-

[4] By R Pintner B V Cunningham and W N Durost. Published by World Book Company, 1938

relation with the Stanford-Binet (1916), the reported coefficients ranging from 73 to .88

Reliability is reported in terms of correlations between the alternate forms (A and B) and in terms of the probable error of test scores The reliability coefficients vary from 83 to 94 The standard deviation of scores is reported as being about five times as large as its probable error, hence this criterion of reliability may be considered satisfactory, since the minimum acceptable relationship between a standard deviation and its probable error is usually given as three to one[5]

## CHICAGO NONVERBAL EXAMINATION[6]

This scale is designed for subjects from six years of age through adulthood,[7] but it is very doubtful whether it is at all adequate for a representative sampling of persons above age thirteen The authors state that the scale has proved clinically useful for children between seven and thirteen years of age In the case of persons above age thirteen, it appears that the tests measure speed of performance

The types of items included are not unique symbol-digit, perception of similarities and classification of objects (by crossing out the one object that is of a different class), three-dimensional visualization and block counting, "paper form board" (marking the parts that would make up a whole geometric figure), visual perception of detail (matching geometric designs which have more or less internal detail), picture arrangement (numbering parts of wholes, to indicate how the parts must be placed to form the whole), numbering pictures in their correct order to represent a logical sequence of events, picture absurdities (noting the missing or superfluous parts), picture matching by relating a part to a whole picture, symbol digit (more complex and more extensive than the first symbol-digit test)

**Validity and Reliability.** The reliability coefficients reported are between 80 and 90 Some of these coefficients are within the range usually regarded as satisfactory But they were calculated for age

[5] The probable error of the standard deviation indicates the extent of fluctuations in scores to be expected as a result of random errors of measurement.

[6] By A W Brown. Published by The Psychological Corporation 1936

[7] The scale may be administered by using verbal directions or by means of pantomime Pantomime, however, cannot be used successfully with children younger than eight years

ranges of two and three years and extend over two to six school grades, whereas reliability coefficients for each chronological age group and each school grade would be more valuable, if a child's performance is to be compared with others of his own age or grade.

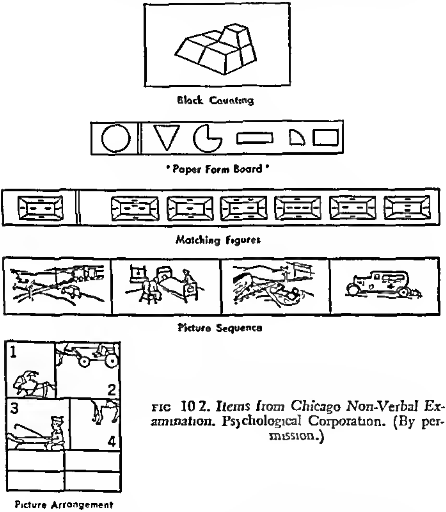



FIG 10 2. *Items from Chicago Non-Verbal Examination.* Psychological Corporation. (By permission.)

The criteria of validity used in standardizing this scale were the usual ones: (1) correlation with CA, (2) known groups (in this instance, the mentally retarded), (3) symmetrical bell-shaped distribution, and (4) correlation with results of other tests.

The reported coefficients of correlation for test scores and chrono-

logical age varied from 57 to 81, using groups having an age range of seven or eight years

Employing the second criterion, the authors of the scale found an average IQ of 61 (S D = 12) for 99 feeble minded children, as com pared with an average of 62 (S D = 6) obtained with the Stanford Binet The mean difference between the Stanford-Binet and the Chicago Nonverbal ratings, however, was 9 0 points (disregarding signs, correlation coefficients were not given) Thus, while the two means are practically identical, the standard deviations and the mean of the differences are such as to indicate that in many individual instances there were significant discrepancies between ratings obtained on the two scales These discrepancies are most probably attributable to the differences in content of the scales, and they suggest, further, that the two scales may be used to supplement each other, rather than as substitutes, when an individual s mental abilities are being analyzed and evaluated

In respect to the third criterion, the authors report a satisfactory distribution of scores, closely approximating the symmetrical curve Like Terman and many other psychologists, they believe that measured intelligence should be and is naturally distributed in the form of the "normal frequency surface ', hence they use that curve as a criterion

Interestingly enough, excepting the mentally deficient, the results of the Chicago Nonverbal Examination have been compared with re sults of two other group tests that are largely verbal in content (Otis and Kuhlmann-Anderson), but not with those of the Stanford Binet which, by most specialists is considered to be a most valuable criterion in standardizing group tests With chronological age not held constant the coefficients fell between 57 and 74, but with CA con stant, the coefficient was only 51 These correlations represent only modern correspondence

## REVISED ARMY BETA EXAMINATION[8]

The original Army Beta scale was constructed in order to provide an examination for illiterates and non English speaking men in the Army of World War I It included the familiar mazes, cube analysis and counting symbol-digit combinations pictorial comple tion, and geometric construction There are two other subtests with

[8] By C. E Kellogg and N W Morton Published by The Psychological Cor poration, 1935

which the reader may not yet be familiar the first is the X O series which includes a number of arrangements of the letters X and O each series has a number of blanks that are to be filled in according to the arrangement of the given sequence For example

| X | X | O | X | X | O | X | X | O | X | X | O | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

The other is a number checking test which consists of a series of paired numbers from short to long The subject is required to check those pairs which are identical For example

| | |
|---|---|
| 650 | 650 |
| 659012534 | 659021354 |

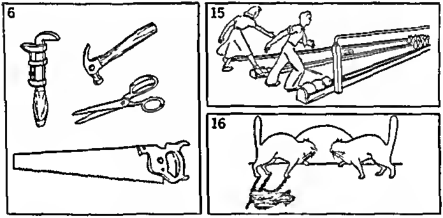



FIG 10 3 *Items from Revised Army Beta Examination* Psychological Corporation (By permission ) For No 6 *the testee marks the* picture absurdity *for Nos 15 and 16 he indicates the* missing parts

The Revised Beta Examination also is intended to serve as a measure of general ability in the case of relatively illiterate or non English speaking subjects The types of subtests of the two scales—original and revised—are not identical for in the revised scale the cube counting and the X O tests have been omitted while picture absurdities and object similarities and differences have been added

Norms are provided in terms of Stanford Binet (1916) mental age equivalents the range being from MA 6 years and 3 months to 16 years and 8 months

**Validity and Reliability.** The scale's reliability coefficients were high when odd-even items were correlated, $r$ being 987. But the coefficient for test-retest scores was 77

Validity of the revised Beta scale is given in terms of the correlation between its point scores and Stanford-Binet (1916) mental ages, the coefficient being $78 \pm 02$ When the revised Beta point scores were correlated with the Otis Self-Administering Tests of Mental Ability (higher examination A), the coefficient found was $.71 \pm 02$ Since these coefficients are based upon results obtained with groups of subjects whose ages varied, the correlations must be interpreted with due regard for the fact that they would be lower if chronological age had been held constant What this signifies, then, is that this nonverbal scale may be used to supplement verbal scales, but it should not be used as a substitute for them

## PINTNER NONLANGUAGE SERIES: INTERMEDIATE TEST[9]

This scale, devised for use from grade 4 through grade 9, differs from most other group scales of general ability in that it utilizes no verbal situations and is independent of word knowledge and language facility except as these are involved in understanding directions The author also provides directions for administering the scale in pantomime where this is desirable, as in the case of subjects who suffer from language handicaps or from defective hearing

The tests consist entirely of materials of a diagrammatic nature, intended to provide "relatively independent" measures of the "spatial factor," "perceptual ability" (visual), and "reasoning" (without use of language) This Pintner scale, therefore, is one of the few which specifically utilize some of the "factors" presumably isolated by those adhering to the group-factor theory of intelligence Yet, interestingly enough, though tables for separate standard score ratings are provided for each of the subtests, no attempt is made to indicate which of the presumed factors is being measured by means of each of the six subtests In fact, the author states that, "No claim is made that the subtests tap primary or independent abilities, and little is known as to the significance of the separate subtest scores . . Only large deviations [from the median] should be given any credence in guidance of individuals "[10] Apparently, then, the inference that this

[9] By R. Pintner Published by World Book Company, 1945

[10] *Manual of Directions* pages 1 and 9

scale measures the spatial factor, perceptual ability, and reasoning rests entirely on an *a priori* basis rather than upon scientific analysis

The total scores of the scale, however, are regarded as having validity, for tables of norms, from age 7 to age 17 are provided, from which mental ages, intelligence quotients, and percentile ranks may be obtained

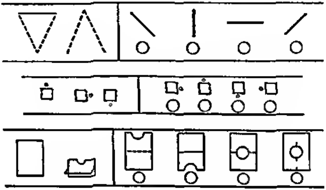


FIG 10 4 *Items from Pintner General Ability Test—Intermediate* Copyright 1938 *by* World Book Company (Reproduced by special permission ) *At top are* Reverse drawings *after looking at the sample pair to the left of the bar, the subject looks at the first drawing to the right of the bar and finds its reverse among the next three drawings For the right middle series* Movement sequence *he finds which of the four drawings at the right completes the series started at the left of the bar For the bottom series,* Paper folding *he indicates which of the items at the right shows the way the folded paper at the left of the bar would look if it were opened up*

**Validity and Reliability.** Reliability of the scale is reported in the usual terms of correlation coefficients between scores on odd numbered items and scores on even numbered items, these being 85 (Form K) and 89 (Form L)

As criteria of validity of his nonlanguage scale, Pintner used increase of mean scores in successive age groups, correlation with his verbal tests of ability, and similarity (judged by observation) between his tests and those devised by factorial analysts The increases in mean scores are moderate, and the correlation coefficients (second criterion) are in the 60 s, for groups whose age range is only one

year Although these coefficients indicate a significant amount of relationship between the verbal and nonverbal tests they are still far enough removed from unity to warrant again the conclusion that these two types of tests may not be used interchangeably but can be used to provide a fuller understanding of an individual s abilities particularly in those instances where a person is handicapped by language disability or where it is desired to discover the existence of a *specific* type of disability (e g in visual perception and discrimination) of a nonverbal kind in the case of a person who has no language handicap

***NONLANGUAGE MULTI MENTAL TEST*** [11]

This scale (in two forms A and B) employs only a single type of item throughout Each item consists of drawings of five objects four of which belong on the basis of some common relationship while the fifth does not belong The testee has to identify and mark the non belonging item

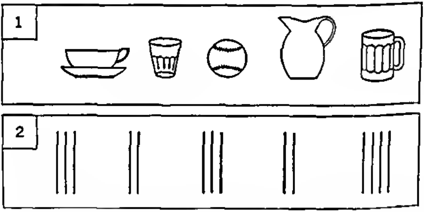



FIG 10 5 *Recognition of Relationships Items from Non Language Multi Mental Test* Teachers College Columbia University (By permission)

The authors state that their scale is constructed to *measure* the ability to recognize and utilize relationships not among verbal symbols but rather among pictorial symbols [12]

Norms are provided for mental age equivalents and grade equiva

[11] By E L Terman W A McCall and I Lorge Published by Bureau of Publications Teachers College Columbia University 1942

[12] *Manual of Directions* p 1

lents in the former instance from a mental age of 33 months to one of 236 months, in the latter instance, from mid first grade to mid-eighth

**Validity and Reliability** Two common criteria of validity were used correlation with chronological ages ($r = 65$) and correlation with mental ages as derived from a variety of group tests of intelligence ($r = 67$)

Reliability coefficients of the scales for a grade range of 3 to 8, were 86 for Form A 90 for Form B and 94 for both forms com bined Reliability coefficients for each of the several grades taken separately varied from 66 to 74 for each form taken alone When both forms were combined, estimated reliability was 80

## PATTERN PERCEPTION TEST[13]

This test, like the one immediately preceding employs a single type of material But it is much more complex its solutions require more subtle perceptions and reasoning and it is intended for use only with adults

The scale includes sixty four items (or problems) Each item consists of a row of five designs the problem being to discern the four designs which form a pattern and to cross out the extra or inappropriate one There are eight sets of items each set placed in the order of increasing difficulty In each set of items the problems begin with an elementary presentation of a theme or pattern which is developed with increasing complexity in subsequent items

Although the Pattern Perception Test is still in process of standardization and development it is described here because it represents a type of nonverbal material that may prove to be very valuable for use

[13] Prepared under the direction of L. S Penrose Published by Galton Laboratory University of London 1947 This type of test was originally planned by Penrose and Raven later Raven standardized a form principally for individual testing During World War II a short form for group use was prepared by Raven for use in the British army See L. S Penrose "An Economical Method of Presenting Matrix Intelligence Tests" *British Journal of Medical Psychology* Vol 20 Part 2 1944 pp 144–146 For background of the present test, see L. S Penrose and J C Raven "A New Series of Perceptual Tests Preliminary Communication" *ibid* Vol 16 1936 pp 97–104 J C. Raven "The R E C I Series of Perceptual Tests an Experimental Survey" *ibid* Vol 18 1939 pp 16–34 P E Vernon "Research on Personnel Selection in the Royal Navy and the British Army" *The American Psychologist* Vol. 2 1947, pp 35 51

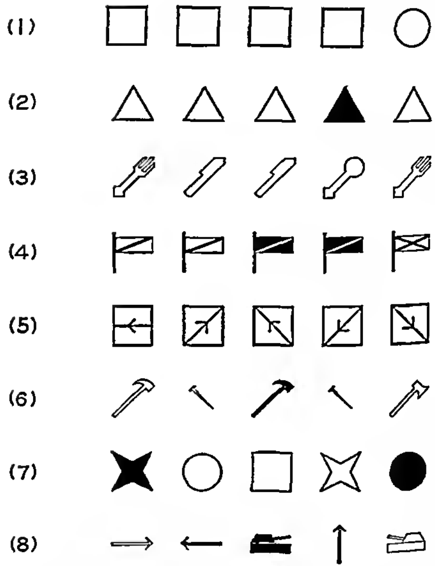



FIG 10 6 *Items from Penrose Pattern Perception Test Galton Laboratory University of London (By permission)*

with a wide range of adult ability This test has been found, for example to have value in identifying men at both the lower and the upper ends of the distribution of mental ability It may prove to be valuable also in examining psychotic persons (or those suspected of

being psychotic), for reports indicate that these persons perform relatively poorly on items requiring judgment (insight) and constructive thinking

Statistical data thus far reported show test-retest reliabilities to vary between approximately 80 and 90 Validity coefficients determined by correlating the test scores with job ratings in the British navy and army varied widely, due to the degrees of reliability of rating criteria and the effects of selectivity in various jobs The *mean* validity coefficients for each of a number of navy and army branches varied from 30 to 47 Correlations with other standardized tests range from a low of 43, for 67 medical students (a fairly homogeneous group) to 73, for a random sample of 597 men in the British army

## PROGRESSIVE MATRICES TESTS [14]

These are nonverbal scales designed to evaluate the subject's ability to apprehend relationships between geometric figures and designs, to perceive the structure of the matrix and of the figure (part) necessary to complete each system of relations (the matrix) presented The tests, thus, are intended to evaluate the person s ability to discern and utilize a logical relationship presented by these nonverbal materials The problems require, in varying degrees, analytical and integrating operations of the kind called insight through visual survey ' Verbalization and abstraction of relationships are also possible factors, if the subject is able to analyze and synthesize by these means Factorial analysis suggests that the matrices tests are measures largely of a "general factor, with a small loading of a spatial perception factor Raven, the author of these tests, interprets this factor as being essentially the same as Spearman's eduction of relations and eduction of correlates

There are several sets and editions of the matrices tests, each intended for a specified age group or for limited ability levels Thus, one set is for children between the ages of 3 and 10, and for mental de-

[14] Prepared by J C Raven Published by H K Lewis & Co Ltd London See by Raven Standardization of Progressive Matrices 1938 *British Journal of Medical Psychology* Vol 19 Part I 1941 pp 137–150 "Progressive Matrices" published by The Crichton Royal Dumfries Scotland (1938–1951) Also G Keir "The Progressive Matrices as Applied to School Children *British Journal of Psychology* (Statistical Section) Vol 2 1949, pp 140-150 G A Foulds and J C Raven "An Experimental Survey with Progressive Matrices (1947)" *British Journal of Educational Psychology* Vol 20 1950 pp 104–110

fectives, another may be used with children from age 6 and with adults, a third is devised for use with only the highest quarter of the population

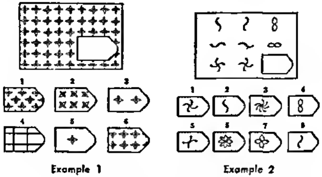



FIG 10 7 *Specimen Items from Raven Progressive Matrices Test* (By permission )

**Validity and Reliability.** While a fair amount of research has been carried out regarding reliability and validity, much more remains to be done in these respects Reliability has been found to differ at the several age levels, and while validity has been suggested in terms of a general factor, the test s predictive efficiency—that is, their practical validity—needs further investigation Even so, this type of instrument is a very valuable addition to available nonverbal tests, particularly since it shows considerable promise for use with adults at the superior as well as average levels Progressive matrices, as a type of test material, and the pattern perception type also, merit intensive experimental study, in themselves, and experimental comparison with the types of materials commonly included in nonverbal tests developed and widely used in the United States At the present time, both of these instruments have progressed sufficiently to warrant their use as valuable supplements to the scales currently employed in both nonclinical and clinical situations for these tests have been found useful with and interesting to individuals of a wide variety of ages, ability levels, and degrees of stability (and instability)

## CATTELL CULTURE-FREE TEST

This is an attempt to provide a measure of general mental ability free from verbal materials and from "the acquired skills of

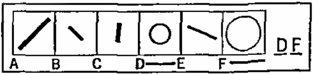
most performance tests "[15] It consists of six parts The first, a classifications test, requires the subject to identify in each row of six figures the two that do not belong with the others In the second, called "pool reflections,' the subject identifies the one of six drawings which represents the specimen drawing as it would appear in a pool image This is a test of spatial perception The third is a completion test in which the testee identifies from among six specimens the one drawing that will complete a series of four members Completions vary from simple matching to the eduction of correlates The fourth, fifth, and sixth parts are called *matrices* In these, the subject has to identify the last member of a series of four or nine parts, the patterns being of increasing complexity Some of the sequences are horizontal, some are vertical, some involve rhythms or cycles The last of the subtests, the sixth, is further complicated by the fact that several sections of the matrix are missing, thus making it more difficult to discern the pattern

**Validity and Reliability.** Reliability is indicated by a split-half coefficient of 88 (corrected by the Spearman-Brown formula), obtained with a group of 121 high-school freshmen

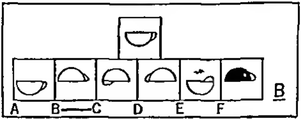
To begin with, this scale was standardized on about one hundred boys in a junior vocational high school and the same number in an academic senior high school In retaining or rejecting an item, the criterion was its ability to discriminate between the two groups, the academic school pupils being regarded as the more able (Note the assumption here that vocational school pupils are mentally less able than the other group The differentials, presumably, are not regarded as due to the age differences ) The retained items were further screened on the basis of responses obtained from college students and from pupils in grades seven and eight Finally, an analysis of items was made on the basis of highest-scoring and lowest-scoring individuals in a group of two hundred students doing major work in psychology

Although this is called a test of general ability, there is a low communality of functions measured, as judged from the intercorrelations of the parts These run from 12 to 63, with a median of 38 The correlations of each of the part scores with total scores varied from 55 to 82, the median being 74 These coefficients, however, are spuriously high because the score of each subtest was included in

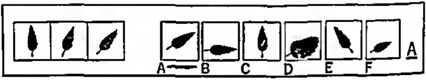
[15] By R B Cattell Published by The Psychological Corporation 1944



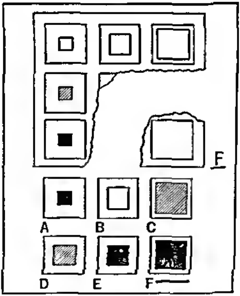

Classification




Pool Reflections




Series




Matrices

FIG 10 8 *Items from Cattell Culture Free Test* Psychological Corporation (By permission )

the total score, the coefficients therefore are in part indicative of self correlation

When results of the Culture Free Test were correlated with scores on the Modified Alpha Examination Form 9, the median coefficients for four groups (N = 376) were about 50 with Alpha verbal and numerical scores taken separately and about 55 with the total Alpha scores The Culture Free Test was also correlated with the Stanford Binet the American Council and the Arthur scales and with the Ferguson Form Board results[16] The mean coefficient with these was 52

The foregoing coefficients would be considered too low as evidence of validity if the other intelligence tests were taken as basic criteria Cattell and his collaborators apparently do not do so they offer these data it seems to indicate the extent of agreement between their test and the others Basically however their criterion is one of factor validity based upon acceptance of the general factor theory and upon factor analysis of their data Their view is essentially this Devise a sampling of tests which have been accepted by psychologists as testing intelligence—in this instance nonverbal perception of relationships and spatial perception that is those tests demanding insight into complex relations ability to learn ability to think abstractly and ability to cope with new situations Analyze the test results to find if they yield a general factor This was done statistical evidence of a general factor was found and Cattell concluded that he and his collaborators were measuring general ability

None of the pragmatic and usual criteria of validity were applied excepting correlations with other tests but to these not much significance was attached This Culture Free Test should be regarded as a tentative and experimental device until its validity is demonstrated in terms other than factorial analysis

## GOODENOUGH DRAWING TEST[17]

This is a test which purports to evaluate a child s intelligence by means of his drawing of a man It is intended for ages 3½ years to 13½ The child is instructed to make a picture of a man as best he

[16] R B Cattell S N Feingold and S B Sarason "A Culture Free Test of Intelligence II Evaluation of Cultural Influence on Test Performance" *Journal of Educational Psychology* Vol 32 1941 pp 81 100

[17] F Goodenough *Measurement of Intelligence by Drawings* Yonkers N Y World Book 1926

ean He is told to work carefully and to take his time Scoring is based not upon esthetic quality but rather upon the presence of essential details which presumably indicate the individual s level of perceptual differentiation of an object that is very familiar in his environment

Although the reliability coefficients are often satisfactory ranging from about 75 to about 90 the drawing test does not correlate well with the other types of tests which have been found most useful and promising in the measurement and evaluation of general ability Yet, this test is another instance of a device which in spite of its doubtful general validity as a test of intelligence has been found useful as an adjunct to verbal tests when mental deficiency is suspected in the case of a given child

## DAVIS-EELLS TEST OF GENERAL INTELLIGENCE[18]

Rationale This test of problem solving ability" is an important innovation in the testing of intelligence It is unusual in respect to its content and the rationale thereof and in respect to its frank rejection of statistical validation with other tests In place of such validation the authors substituted intensive interviews with children to disclose the mental problems of a kind found in most of the basic areas of children s lives school home play stories and work The specific problems resulted from intensive observation and detailed *interviewing of children in many areas of activity* The extent to which the items in the test deal with problem situations which seem real to children of the age levels for which the test is planned may best be judged by examining the test with this criterion in mind The test items selected for initial tryouts and experimentation were based upon the insights of a number of educators and sociologists familiar with the characteristic modes of living and child upbringing at different socio-economic levels and in part upon systematic observation of children in free time activities (on playgrounds in neighborhood groups in schools etc ) [19]

Contents The instrument finally emerged after extensive research and six tryouts with large numbers of children from widely different socio-economic backgrounds in several sections of the United States It is believed to be culturally fair to all socio-economic groups in *urban* areas

[18] By Allison Davis and Kenneth Eells. Published by World Book Co., 1953
[19] *Manual* pp 7 and 18

Forms for two levels are available Primary for grades 1 and 2, Elementary for grades 3 through 6 The items finally selected are 88 in number, as follows

*Best Way items* (29) In each item three pictures show the beginnings of attempts to solve a stated problem or perform a given

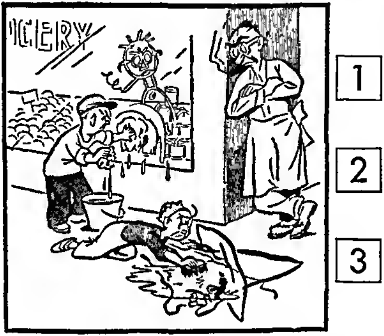



FIG 10 9 *The task is to select the one statement that best explains the situation shown in the picture*

1 *The boys want to wash the man s window and sidewalk*
2 *The man is making the boys wash his window and sidewalk*
3 *You can t tell from this picture why the boys are washing the window and sidewalk*

From Davis and Eells Test of General Intelligence World Book Co (By permission )

task The subject indicates the picture that will lead to the best solution of the problem

*Probabilities items* (29) Each picture shows a situation in which certain elements are present The child has to select from three choices the most probable explanation of what is happening in the picture

can He is told to work carefully and to take his time Scoring is based not upon esthetic quality but, rather, upon the presence of essential details which presumably indicate the individual's level of perceptual differentiation of an object that is very familiar in his environment

Although the reliability coefficients are often satisfactory, ranging from about 75 to about 90, the drawing test does not correlate well with the other types of tests which have been found most useful and promising in the measurement and evaluation of general ability Yet, this test is another instance of a device which, in spite of its doubtful general validity as a test of intelligence, has been found useful as an adjunct to verbal tests when mental deficiency is suspected in the case of a given child

## DAVIS-EELLS TEST OF GENERAL INTELLIGENCE[18]

**Rationale.** This test of "problem solving ability" is an important innovation in the testing of intelligence It is unusual in respect to its content and the rationale thereof, and in respect to its frank rejection of statistical validation with other tests In place of such validation, the authors substituted *intensive interviews* with children to disclose the "mental problems of a kind found in most of the basic areas of children's lives school, home, play, stories, and work The specific problems resulted from intensive observation and detailed interviewing of children in many areas of activity The extent to which the items in the test deal with problem situations which seem real to children of the age levels for which the test is planned may best be judged by examining the test with this criterion in mind " The test items selected for initial tryouts and experimentation were based upon the 'insights of a number of educators and sociologists familiar with the characteristic modes of living and child upbringing at different socio-economic levels, and in part upon systematic observation of children in free-time activities (on playgrounds, in neighborhood groups, in schools, etc ) "[19]

**Contents.** The instrument finally emerged after extensive research and six tryouts with large numbers of children, from widely different socio-economic backgrounds, in several sections of the United States It is believed to be culturally fair to all socio-economic groups in *urban* areas

[18] By Allison Davis and Kenneth Eells Published by World Book Co , 1953
[19] *Manual* pp 7 and 18

Forms for two levels are available Primary for grades 1 and 2 Elementary for grades 3 through 6 The items finally selected are 88 in number as follows

*Best Way items* (29) In each item three pictures show the beginnings of attempts to solve a stated problem or perform a given

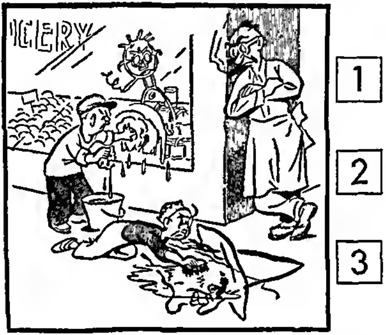



FIG 10 9 *The task is to select the one statement that best explains the situation shown in the picture*

1 *The boys want to wash the man s window and sidewalk*
2 *The man is making the boys wash his window and sidewalk*
3 *You can t tell from this picture why the boys are washing the window and sidewalk*

From Davis and Eells Test of General Intelligence World *Book Co (By permission )*

task The subject indicates the picture that will lead to the best solution of the problem

*Probabilities items* (29) Each picture shows a situation in which certain elements are present The child has to select from three choices, the most probable explanation of what is happening in the picture

*Picture Analogy items* (22) This is the familiar type of item in which a relationship between two objects is shown, and the subject is required to find a similar relationship in a given set of pictures

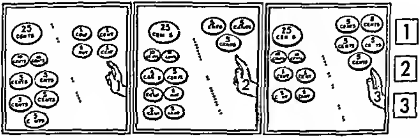


FIG 10 10 *The task is to select the picture in which a given sum of money can be made beginning with the right hand side of the dotted line and completing from the left side In this problem, the sum to be made is 40 cents* From Davis and Eells *op cit* World Book Co (By permission )

*Money items* (8) In each item, two sets of coins are shown in three different combinations Each combination is incomplete The problem is to discern the appropriate combination, of the three, that will yield a stated sum when completed

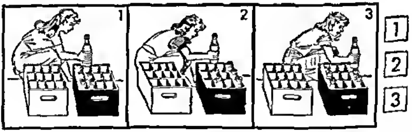


FIG 10 11 *The task is to place the bottles in the black box in such a way that the white box may best be placed on the black one* From Davis and Eells, *op cit* World Book Co (By permission )

Of these, 26 are in the Primary form only, 41 are in the Elementary only, and 21 are common to both.

**Validity and Reliability.** Although Davis and Eells do not base the validity of their tests upon correlations with other and earlier scales, they do present, for informational purposes, some correlations with

the Otis Quick-Scoring tests, obtained for groups in single grades, 3 through 6 Of the sixteen coefficients, seven are in the 50's, the remainder are rather evenly distributed from 39 (lowest) to 66 (highest) The authors believe these coefficients are what should be expected, since they indicate that the abilities measured by their tests bear a substantial relationship to those measured by the other tests, yet their tests and the others do not measure altogether the same factors

Correlations with standardized school achievement tests are also reported These are

| | |
|---|---|
| with reading | 43 |
| with arithmetic | 41 |
| with language | 40 |
| with spelling | 24 |

These coefficients are significantly lower than those found with the more usual types of individual and group scales, but this is to be expected because of the very nature of the Davis-Eells tests, and since the authors' position is that several significant factors other than problem-solving ability contribute to success in school achievement

Split-half reliability coefficients range as shown below

| | |
|---|---|
| grade 1 | 68 |
| grade 2 | 82 |
| grade 3 | 84 |
| grade 4 | 83 |
| grade 5 | 82 |
| grade 6 | 81 |

A coefficient of 68 is much too low for predictive purposes The remaining indexes are moderate, but not as high as is optimally desirable

Test retest reliabilities (two week interval) are

| | |
|---|---|
| grade 2 | 72 |
| grade 4 | 90 |

On the whole, these coefficients, if typical, indicate that the present test is not so reliable for refined differentiation as it is for differentiating between levels, for example, very superior, superior, average, inferior, very inferior

In connection with the validity and general characteristics of these tests, several aspects should be pointed out The authors aimed to develop an instrument that would present situations and problems in forms that are within the experiences common to all children in the groups for whom they were planned This meant the elimination of language (except *in the directions*) and other cultural factors that might favor one group and handicap another While nonverbal tests are not new, the kinds of situations presented are novel Furthermore, the psychological functions to be sampled by these and similar problem situations were determined and specified after psychological interviews and psychological analysis, rather than by factorial analysis

**Evaluation.** *The foregoing difference in procedure is very significant* with regard to principles of test construction Factorial analysts devise a large number of test items and subtests on the basis of certain assumed mental processes involved, the scores on these are intercorrelated and further analyzed in order to find common functions to combine and reduce the number of subtest categories, and to name them Davis and Eells, on the other hand, first ascertained the kinds and range of problems children deal with then by *interview* and analysis of children s responses they determined that certain psychological functions were operative and had differentiating significance The functions they specify are association insight, reasoning, and organizational ability (method of attacking a mental problem) After the items for these tests were developed, each of a group of children was interviewed to determine whether the problems evoked the mental processes which the tests seek to measure It was found that 92 percent of the children who answered analogy problems correctly also explained the analogy relationship correctly And nearly all pupils explained the relationships correctly in solving the other types of subtest problems These findings strongly indicate that the test problems evoked in the successful subjects the mental processes which the authors intended should be utilized

Like all other new tests, the Davis-Eells will have to be subjected to investigation to determine their predictive efficiency Since the authors point out that school achievement is dependent only in part upon problem-solving ability (which is to be measured by their tests), *it is to be expected that predictive efficiency will not be very high*, in terms of correlations, with school grades It will be necessary, *then*, to

use other selected criteria of demonstrated problem-solving ability, or to utilize pupils in whom this ability is the only significant variable, while the others are practically constant

## EVALUATION OF NONVERBAL GROUP SCALES

Uses. A survey of available nonverbal scales shows, for the most part, that they are valuable with children who have had limited educational opportunities or impoverished social backgrounds, with young children who have not yet learned to read, with older pupils who are handicapped by reading or language difficulties, and with illiterate or non-English-speaking adults Possible exceptions to this statement of limited usefulness are the Pattern Perception, the Progressive Matrices, and the Davis Eells Tests

Nonverbal tests are valuable, also, for the better diagnosis of cases who, on verbal tests, have intelligence quotients between about 60 and 75, and who, therefore, would be considered as subjects for special educational treatment or possibly institutional care The examining psychologist might be in doubt with regard to such borderline cases, but if the results of the nonverbal tests confirm those of the verbal, he has reason to allay his doubts However, if the rating on the nonverbal tests is significantly higher, then the case will require further study to account for the discrepancy

Nonverbal tests can be clinically useful, also, with individuals whose intelligence quotients are higher than 75, that is, for individuals who, on verbal tests, *appear* to be significantly less capable than there is reason to believe they actually are, on the basis of other information about them In this connection, they are particularly useful in population centers having large numbers of non-English-speaking homes

In whatever situation nonverbal tests are used, the examiner must realize that defective vision or slow psychomotor responses can be a handicap The first of these handicaps points up the importance of clear drawings—a condition that is not always satisfied

Tests of mental ability have had their greatest usefulness in schools, where they have been utilized for purposes of educational and vocational guidance, as well as in the diagnosis of learning difficulties in the case of a particular individual Nonverbal group tests have been found valuable in efforts to determine aptitude and promise in shop work, mechanical drawing, architectural drafting, and occupations of a mechanical or quasi-mechanical nature—all of which make demands

upon those psychological activities which enter into problems involving geometric perceptions and reasoning with the concrete rather than with the abstract

For these purposes nonverbal group scales may not be used indiscriminately for the following reasons Some are measures principally of detailed visual perception (Revised Army Beta) some attempt to span too wide an age range and are not able to make any but rather gross differentiations (Non language Multi Mental Test which is scaled for ages 33 months to 236 months) at the higher age levels some are less suitable for women than for men because of specialized content

**Validity** Studies of the validity of nonverbal scales show that while most of them correlate significantly with scales of the verbal type (individual and group) the coefficients are far enough removed from unity to warrant using the two types as supplements rather than as equivalents When scores on verbal and on nonverbal scales are correlated for children in the earlier grades (approximately through grade six) the coefficients obtained are usually in the 60 s and 70 s with relatively few in the 80 s But when the subjects tested are pupils in the later grades the coefficients usually fall in the 50 s and 40 s with a few lower and a few higher These generally lower coefficients in the case of pupils in the later grades are due to the inability of most available nonverbal tests to discriminate between individuals in the upper levels of ability

**Functions Measured.** The reader will recall that most though not all, authors of nonverbal tests of mental ability seek to measure the same mental processes as those tested by means of verbal scales Some of these authors are unequivocal in maintaining that the nonverbal tests require essentially the same type of intelligent performance as that required by the abstract symbols of language and number They hold that the problems presented in diagrams pictures charts and geometric forms closely parallel those presented by means of language and number For example picture arrangement is regarded as being similar in function to disarranged sentences picture analogies similar to word analogies picture completion similar to sentence completion reasoning with geometric patterns similar to reasoning with numbers and words perceiving similarities differences and part

whole relationships in pictures and patterns similar to such relation ships in language Many nonverbal tests however suffer from the fact that they attempt to assess general ability by means of homogeneous test items or by means of a very limited number of subtests

Related to the question of measuring general ability by means of a more or less homogeneous nonverbal scale, it is important to note that it is quite possible to find *some sort* of general factor—which is actually a statistical and theoretical concept—in a series of subtests such as those in Cattell s and others but the statistical derivation of a general factor does not in itself give proof or assurance that the tests are necessarily measuring the functions being attributed to them For example a series of tests might be devised in which a "general factor" emerges and which upon *psychological* analysis proves to be a factor of speed of work In another instance a general factor of visual acuity may emerge etc Fundamentally in any statistical analysis of test data the factor or factors that emerge will depend upon the initial characteristics of the tests themselves In establishing validity of a scale, therefore it is inescapably necessary to analyze the psychological processes involved in the tests and to compare the results they yield with forms of activity and evidences of ability which are indicative of intelligent behavior in our culture

The significant coefficients of correlation found between verbal and nonverbal tests of intelligence demonstrate that there is merit in the view that the two types are in a significant degree measuring the same or associated functions But this does not mean that verbal and nonverbal tests are equivalent for one type may also involve certain functions not involved in the other or one may demand a higher level of the same functions being tested than does the other

Language and number are symbolic systems which *represent* something else e g , objects qualities events actions Development of abilities in language and number facilitates intelligent behavior, for the use of these symbols expands the individual s range of experience beyond the limits of the immediate situation Development of language and number makes possible a finer discernment of forms and objects in the world surrounding the individual, for with the use of language and number he is enabled to analyze, synthesize, classify, and organize his perceptions Objects and events, at first vague, are more sharply defined, likenesses and differences are accentuated, evaluations are refined Language and number enable individuals

also to organize their thinking into larger and more comprehensive unified patterns

It is because the use of language and number requires that the individual go beyond the immediate concrete situation and because he thereby can engage in more complex and subtle mental operations that many psychologists regard ability to deal with symbols as a higher form of intellectual activity than the ability to deal with concrete objects They therefore prefer to test intelligence, whenever possible and appropriate, by means of verbal and numerical materials They would, however, use nonverbal tests, when these are made necessary by developmental immaturity, language, or cultural handicap, to gain the insights that these tests provide if they are adequately scaled in difficulty

**Cultural Influences.** The emphasis upon verbal and quantitative aspects of intelligence in many of the individual and group scales has given rise to a misapprehension regarding the nonverbal scales namely, that these latter are "culture free", and, in fact, one test author (Cattell) has so named his scale Inspection of the items in this scale and others reveals that they utilize many objects that children and older persons learn about through experiences in their environments These experiences are dependent upon a culture just as development of verbal and quantitative abilities are. The differences are matters of degree of cultural influence and universality or near-universality of experience Consequently, it is preferable to speak of "culture fair" tests rather than "culture free" tests, in connection with tests which utilize materials that do not handicap or favor any segment of the population for whom the test is intended

The presence of cultural influence in a test that *appears* to be "culture free" was demonstrated in a study made with several tribes of North American Indians[20] The Goodenough Draw-a-Man Test was used In a group of Hopi Indians, the mean IQ for boys was 123, while for girls it was 102 Zias also showed appreciable differences in favor of boys, whereas in a group of Navahos, the means of boys and of girls were very nearly equal (107 and 110) The sex differences or similarities within each tribe are attributed to sex differences or simi-

[20] R. J Havighurst et al, "Environment and the Draw a Man Test the Performance of Indian Children," *Journal of Abnormal and Social Psychology Vol 41, 1946, pp 50-63*

larities in training and experience within each culture Boys and girls are trained to observe different aspects and details of their environment and are taught different types of drawing The two sexes have different functions in their group, these functions are reflected in differentiated training, the differences in training are reflected in differences in performance

Since every person must develop in an environment of some kind, his skills, information, repertory of responses, modes of thinking, etc , are to some extent culturally conditioned Some psychological tests are more "culture fair" than others At this point we recall again Binet's principle that a test of intelligence should be consonant with the milieu of those who are to be measured by it

# II.

## VERBAL AND MIXED GROUP SCALES OF MENTAL ABILITY

THE scales presented in this chapter are either entirely or predominantly verbal in content The proportion of nonverbal materials varies in different scales, but it will be noted that even in those instruments that contain appreciable portions, symbolic materials (language and number) predominate

Since there are many scales that come within this classification, it is neither possible nor necessary to describe and analyze all of them It is the purpose of this chapter to present a sufficient number of representative scales so that the student can know their characteristics and content, their quality, advantages and disadvantages, their similarities and differences The details of content are presented in order to provide the student with a clearer conception of the psychological processes being tested than would otherwise be the case The statistical details for each test will enable the student to see clearly the techniques of standardization used and the degree of success achieved, upon which evaluations of these instruments must rest

### CALIFORNIA TESTS OF MENTAL MATURITY[1]

**Contents.** The revised series provides scales on *five levels* preprimary (kindergarten and entering grade 1); primary (grades 1–3), elementary (grades 4–8), intermediate (grades 7–10 and

[1] By E T Sullivan W W Clark, and E W Tiegs. Published by California Test Bureau, Los Angeles 1951

adult), advanced (grade 9 and adult) All of these are designed to test the same "mental factors', hence, since the series covers the wide age range indicated, it is essential that the content of each scale be adapted to its particular level, as regards difficulty and form Thus at the earliest levels there must be emphasis upon nonverbal materials, with minimum requirements made upon word knowledge and number concepts The later levels increase in their demands upon word knowledge, number concepts and reading of and reasoning with numerical and verbal materials, while increasingly complex nonverbal materials are also retained

We shall describe only the elementary scale (grades 4–8), since it is representative of the entire series

The elementary scale consists of twelve subtests, grouped under five headings, or factors, as follows

*Memory*—

1 Immediate recall series of words pronounced in pairs, then only the first word of each pair is repeated and the subject is to recall the second word of the pair

2 Delayed recall a story is read to the subjects Thirty minutes later they are given a series of multiple choice items to test the extent to which details of the story are recalled

*Spatial relationships*—

3 Sensing right and left 20 pictures of hands and feet in various positions The task is to discriminate between right and left

4 Manipulation of areas spatial patterns of a variety of forms and in different positions to be manipulated, to test spatial imagery

*Logical reasoning*—

5 Opposites 15 sets of drawings showing five objects in each set The first is the usual 'given object, the testee selects from the other four the one that is the opposite of the first

6 Similarities the well-known classification test, using 15 sets of drawings The first three in each set are alike in some respect The task is to select the similar item in the remaining four drawings

7 Analogies the familiar test of relationships, using drawings The first two stand in some relationship (a hat and a man's head), a third item is given (a shoe) The task is to select a fourth item

that bears the same relationship to the third as the first does to the seeond

8 Ioference a major and a minor premise are given The task is to select a logieal cooelusion from among several alternatives

*Numerical reasoning—*

9 Number series each series of numbers increases or decreases according to a principle The testee has to discern that principle

10 Numerical quantity number concepts using coins The subject indicates how many coins of each denomination are required to make up a specified sum

11 Numerical quantity arithmetical problems

*Verbal concepts—*

12 Word similarities 100 given words In each item, the task is to select one word, from four, that is synonymous, or nearly so, with the given word

This scale also provides three optional subtests of visual acuity, auditory acuity, and motor coordination These are not ineluded in the scoring They may be used to learn whether, in respect to these funetions the subject will or will not be handicapped on the subtests that are seored The use of these three preliminary tests ean be very helpful in identifying persons who would be under a disadvantage in taking a group test, and who should therefore, be examined individually When an individual scale is used, existing handicaps ean be overcome, in part at least, and can be taken into account in the interpretation of the testee's performance and score

Scoring An aspect of these scales worthy of special note is the method of scoring The score is obtained for each of the divisions (e g, immediate recall, delayed recall) under each of the factors (memory, etc ) For each separate score, a mental age rank is found and plotted on a scale, or profile The separate scores of each division are then added to give the score for the particular factor (in this instance, memory), which is also plotted on the profile The scores of all factors are then added to yield the total score, from which the usual indices may be derived Also, the appropriate subtest scores are added to obtain a rating on language factors (subtests 2, 8, 11, 12),

then the scores on the remaining subtests are added to get a rating on nonlanguage factors

Thus these scales enable the examiner to obtain (1) ratings on the several subtests (2) separate mental ages and intelligence quotients for verbal subtests combined nonverbal subtests combined and total score of all subtests This type of scoring and profile permits ready analysis of the subject s weaknesses and strengths consistencies and inconsistencies in the types of mental operations being tested as suming of course that the test materials are valid and the reliability is high

**Validity and Reliability** Although the norms for this scale are based on a controlled (stratified) sampling of over 125 000 cases [2] reliability statistics were based on only 725 pupils in grades 4 to 6 in representative school districts The split half method (presumably odd-even) was used the coefficients having been cor rected by the Spearman Brown formula The results were

total mental factors coefficient = 95
standard error of measurement 3 5 IQ points
language factors coefficient = 94
standard error of measurement 3 9 IQ points
non language factors coefficient = 92
standard error of measurement 4 5 IQ points
87 (spatial relationships) to 92 (memory)
Standard errors of measurement in the subtests range
from 4 5 to 5 8 months of mental age

On the whole these coefficients fall within the range of quite satis- factory correlations particularly for the two major divisions and for the total score The standard errors of measurement in terms of IQ points also compare favorably with those of the sounder tests Al though the errors of measurement for the subtests are somewhat larger *this is what would be expected* *Once again* *greater reliability* of wide rather than narrow sampling of performance is demonstrated

It is in respect to the accepted validity criteria that the authors of these California scales do not provide adequate data [3] The authors

[2] *Manual* p 22 The soc o-econom c d stribut on of the sampling is not re ported in the manual

[3] In a m meographed prospectus kindly furnished by the tests authors there

purpose, to begin with, is to measure most of the kinds of mental processes sampled by the Binet scales, based upon an analysis of the "conceptual framework" of the Binet, but no correlational data with the Stanford Binet are given in the manual. Unpublished data, provided by the publishers of these scales, show the following results (Table 40) obtained with the Elementary scale.

TABLE 40

Correlations between California Elementary Scale and S-B

| California | | Stanford Binet | | | |
|---|---|---|---|---|---|
| Md IQ | S.D IQ | Md IQ | S.D IQ | N | r |
| 107.3 | 28.5 | 108.2 | 29.3 | 283 | 94 |
| 122.0 | 27.9 | 128.5 | 28.8 | 182 | 93 |
| 79.2 | 22.7 | 79.5 | 21.6 | 101 | 90 |

Several significant facts and inferences emerge from the foregoing table. First, there is close correspondence between the two scales in respect to medians and standard deviations of intelligence quotients. Second, the groups are not representative of the population for which the scale is intended, since the medians are significantly above or below 100 and the S.D.'s are very much larger than would be found in an unselected population (namely, about 16). Third, due to the very wide range of ability, as represented by the very large standard deviations, the probability is that the correlation coefficients are significantly higher than they would be for a population of narrower and typical range. Fourth, it would be desirable to determine this aspect of validity separately for each age group, since close correspondence at one age level or for combined age levels does not necessarily assure equally close agreement at other age levels.

The manual furnishes data on the intercorrelations of the subtests, based on 1048 cases in grades 4 to 6. In view of the authors' previous preference for the group factor theory of intelligence (stated in earlier editions of the scales) and of the generally low intercorrelations of the

are extensive data dealing with the *1947 short form of the advanced scale. This prospectus reports significant correlations between the California* and other group scales. For the "language factors" of the California scale the coefficients are largely in the 70's, whereas for the "non-language factors" they are in the .50's. Correlation coefficients with school grades vary, as they do with all tests of intelligence, and are higher for the "language factors." For these factors the correlations are, for the most part, in the .50's, 60's, and 70's. Similar data are not provided for the elementary scale described herein.

present subtests, we may conclude that these coefficients are regarded as one evidence of the scale's validity namely, that the subtests have relatively little in common regarding measured factors, hence they satisfy one requirement of the accepted theory [4]

The subtest intercorrelations may be summarized as follows

range of subtest intercorrelations 25 (spatial relationships and numerical reasoning) to 60 (memory and verbal concepts)

seventy percent of the coefficients are below 50, fifty percent are below 40

range of correlations of subtests with language subtests 35–95 (in part self-correlations)

range of correlations of subtests with nonlanguage subtests 55–78 (in part self correlations)

range of subtest correlations with total scores 60–86 (in part self-correlations)

Regarding validity, one may draw some inferences, however, from data provided in other connections (1) Median IQ for the entire population sample is 100, with a standard deviation of 16 (2) The population samplings at the higher educational levels show a progressive increase in medians and decrease in standard deviations, until at the college sophomore level these are 114 5 and 13 5, respectively, and at college graduate level they are 124 and 12 Such progressive increases are to be expected, since advancing educational levels are more or less selective as regards intelligence

Even when a scale is validated within a theoretical conception and framework, it is still necessary to subject the scale to other validating procedures to determine if it actually serves the purposes for which it is intended In other words, does the scale "work'? What is its predictive efficiency? The necessity of answering these questions is the reason why all tests of intelligence should be validated against accepted criteria of intelligent behavior Just devising test materials that satisfy a theoretical conception gives no assurance of the scale's practical and predictive validity

---

[4] The reader will recall that the original Binet scales and the Stanford Binet are based upon a general factor theory

In the test manual page 4, it is stated that "The total mental factors score has been found by the authors and other investigators to correlate as high or higher with the individual Stanford Binet than any other mental ability test" No data relevant to this aspect are given in the manual but the several correlations reported above were provided in a personal communication

## TERMAN-McNEMAR TEST OF MENTAL ABILITY[5]

**Contents.** This scale (in two equivalent forms) is intended for use primarily in grades 7 through 12, though norms are provided from the age of 10 years through 19 years, 11 months The scale consists of seven subtests information synonyms, logical selection, classification, analogies, opposites, and best answer

*Items from*
*Terman McNemar Test of Mental Ability*
(World Book Co By permission)

*Information*

Polo is a kind of
(1) disease (2) work (3) bear (4) game
(5) language

*Synonyms*

Comic—(1) clumsy (2) laughable
(3) universal (4) tricky
(5) peculiar

*Logical selection*

An orchestra always has
(1) violinists (2) piano (3) musicians
(4) saxophone (5) singers

*Classification*

(1) Catholic (2) Methodist (3) Presbyterian
(4) Republican (5) Baptist

*Analogies*

Zoo is to animal as aquarium is to
(1) birds (2) fish (3) bees (4) statues
(5) butterflies

*Opposites*

Exit (1) emit (2) transcend
(3) entrance (4) origin
(5) arrival

*Best answer*

The saying, "Idle brains are the devil s workhouse " means
(1) The devil is lazy
(2) People who are idle get into trouble
(3) Many hands make l ght work
(4) The devil works with his brains

[5] By L. M Terman and Q McNemar Published by World Book Company 1942. This scale is a revision of the Terman Group Test of Mental Ability which was published in 1920

The content of the scale is quite homogeneous in that it is entirely verbal in character. The scale is thus consistent with Terman's definition of intelligence, that is, the ability to deal with symbols and abstractions.

The authors subscribe to the general factor (*g*) theory of intelligence. They hold that the general factor is best tested by means of materials using symbols and abstractions. In order to achieve a high degree of homogeneity in test materials, they even omitted from their scale arithmetical and numerical types of subtests, which are widely regarded as being very good tests of the general factor. The authors state the reason for their selection of materials as follows: "More homogeneous material has been used in order to have a test more highly saturated with a common factor or ability. Thus, the exclusion of arithmetical and numerical subtests means that the scores of any two individuals are more nearly comparable qualitatively, i.e., they lie along the same continuum. This continuum may be characterized as general verbal intelligence."[6] The usefulness of this scale, therefore, is restricted to subjects who are not laboring under a language handicap and to situations wherein 'verbal intelligence' is required as the sole or major ability.

**Validity and Reliability.** Reliability of this scale is presented in terms of three familiar methods: the split half method (correlation of scores on odd-numbered and even numbered items), the interform method (correlation of scores on the two forms given to the same subjects), and the probable error of measurement. The split half reliability coefficient was .96 (279 cases, grades 7 through 9), while the interform reliability coefficient was .95 (239 cases, grades 7 and 9). When an age range of only one year was taken (13–6 to 14–5), both coefficients were .96. The probable error of measurement was found to be 2.2 standard score points. (In terms of the scoring units of the scale, this probable error can be considered small.)

Validity of individual items was determined primarily on the basis of percent of pupils passing each item in the successive grades—in other words, the extent to which each item differentiates between groups at different levels of maturity. A second criterion of item validity was the correlation of each item with total scores.[7] No item

[6] *Manual of Directions*, p. 1.

[7] For this purpose the "tetrachoric correlation" was used. This technique differs from the more familiar one used in correlating two sets of variable

was retained if it yielded an average coefficient of less than 30, in fact, ninety percent gave correlations of 40 or higher, with an average of 53 Thus, for validity, the final selection of items was made on the basis of item difficulty and degree of relationship of item performance to total scale performance

Scoring Two statistical factors are at the base of the rating method used (1) a single origin, and (2) comparable units in all parts of the scale The reader is already familiar with the fact that raw scores do not provide comparable units throughout a scale For the Terman-McNemar scale, this problem is met by a type of standard-score scale which uses the median of the 14-year age group of the national standardization population as the origin, and the standard deviation of this age group, arbitrarily made 16 points as the unit of measurement 'Scores on this scale for all age groups are thus measured from a single origin and provide comparable units throughout all parts of this scale' [8]

In other words, the authors of the scale have devised a method of scaling which is a variant of the familiar standard score The 14 year age group is taken as the standard, norms of other age groups are given in terms of standard scores of the 14-year group Thus, for age 14, the median raw score was 76, this was arbitrarily called a standard score of 100 The raw score standard deviation (actually about 27) was assigned an arbitrary value of 16 No reason is cited for the choice of this particular number It may be noted, however, that 16 was the most typical value of the standard deviation for the IQ s of the 1937 revision of the Stanford Binet and thus is a convenient number to use in a group scale which attempts to measure the same type of ability Having assigned this arbitrary value, the authors then proceeded on the assumption of normal distribution and from a table determined the standard scores that would correspond to various raw scores in a distribution that had a mean of 100 and S D of 16 For example, the average raw score of the 13-year group (13 years 0 months) is 63 In a normal curve with a standard deviation of 16 points and a mean of 100, the score of 63 would have as its equivalent

scores in that the underlying assumptions are different but interpretation of obtained coefficients is approximately the same For an explanation of the tetrachoric method see any standard textbook in statistics

[8] *Manual of Directions* p 4

a standard score value of 93, which then becomes the norm for the age 13 years and 0 months

The procedure for finding an individual s standard score rating on this scale amounts to this the raw score is obtained, each raw score is found in a table which gives its standard score equivalent, eacb standard score is found in another table giving its equilavent mental age The IQ, however, is not found by means of the usual formula (IQ = MA/CA) In fact, the mental age is not used in calculating IQ for this scale Instead, the authors provide a table for finding what is called the 'deviation IQ,' so called because 'Basically the procedure for computing deviation IQ s requires that the difference be found between the [individual s] obtained standard score and the average standard score for other individuals of the same age This difference or deviation is then interpreted directly in terms of IQ [from a table] This can be done because both IQ s and the normalized standard scores are distributed normally "[9] In other words, Terman and McNemar assume that raw scores and IQ s are both normally distributed When this assumption is made, it is possible through knowledge of the characteristics of the two curves, to transmute one set of scores directly into the other

We have presented this scale's method of deriving IQ s in some detail because it is essential that the student of psychological testing be aware of the several techniques and of the fact that indexes called by the same name are not always derived in the same manner First, we have the original and most common method of finding IQ mental age divided by chronological age Second, there is the method used with the Bellevue scale And third, there is the type of deviation IQ described above

## TESTS OF PRIMARY MENTAL ABILITIES[10]

**Contents.** The Thurstones have published a series of group tests variously known as The Chicago Tests of Primary Mental Abilities and The SRA Primary Mental Abilities The latter are more recent scales, shorter and less satisfactorily standardized and reported than the original Chicago PMA tests, and on the whole inferior to the earlier versions The following description, therefore, will be limited

[9] *Manual of Directions* p 9

[10] By L. L. and T G Thurstone Published by Science Research Associates several forms, 1938–1950

to the Chicago PMA tests (1943), devised for ages 11–17 This particular scale will serve our purpose, since all of the scales in the Thurstones' series are based upon the same psychological and statistical principles, even though all are not of equal merit

The PMA scale for ages 11 to 17 is constructed upon the group-factor theory of mental ability, that is, upon the theory that intelligence consists of the operations of certain distinguishable and relatively independent mental functions (See Chapter 3 ) The "primary abilities" to be tested by means of this scale are those which L L Thurstone and his collaborators report as having been isolated by factorial analysis The "primary factors" measured are six number facility, verbal comprehension, spatial perception, word fluency (extent of word associations as distinguished from verbal comprehension), reasoning, and rote memory Each of these is measured as indicated below

*Number facility* by tests of addition and multiplication

*Verbal meaning* by one test of vocabulary (the familiar multiple-choice form) and one of supplying words to fit given definitions In each item of the latter, five letters are provided, one of which is the first letter of the correct word

*Spatial perception*, by tests in which designs and geometric figures, differently rotated, are to be identified as being the same as or different from a given design or figure

*Word fluency*, by two tests—one requiring that, within a time limit, as many words as possible be written, beginning with a given letter; the other requires that as many four letter words as possible, beginning with another letter, be written within a time limit

*Reasoning* by two tests involving, in one, perception of the patterns within series of letters of the alphabet, and, in the other, perception of the patterns within letter groupings

*Rote memory*, a names test There are twenty cards, on each of which are a first and last name The cards are exposed consecutively, each for fifteen seconds The subjects are then required to pair off each last name with its correct first name (chosen from seven names in multiple-choice form).

**Validity and Reliability.** By means of the split half method, reliability coefficients were separately calculated for five of the six subtests for grades 6, 8, 10, and 12, with approximately 200 subjects at each half grade level No reliability coefficients are provided for "word fluency," since the scores on this subtest do not lend themselves to the split-half method Instead, it would be necessary to use the re test

## COMPLETION

Read the definition below. Think of the word which fits the definition. The first letter of the word is in the row of letters under the definition.

| |
|---|
| The first meal of the day.<br>A = B C = D = E = |

The word is "Breakfast." "B" is marked because it is the first letter of the word "Breakfast."

Do the following example.

| |
|---|
| A place or building for athletic exercises.<br>D = G = H = T = V = |

## FIGURES

Look at the row of figures below. The first figure is like the letter F which is right side up. All the other figures are like the first but they have been turned in different directions.

Satisfy yourself that all of these figures look like the first one if they are turned right side up.

Now look at the next row of figures. The first one looks like an F. But none of the other figures would look like an F even if they were turned right side up. They are all made backward.

Some of the figures in the next row are like the first figure. Some are made backward. The figures like the first figure are marked.

A B C D E F

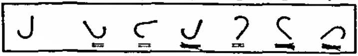


FIG. 11.1 *Items from Chicago Tests of Primary Mental Abilities.* Science Research Associates. (By permission.)

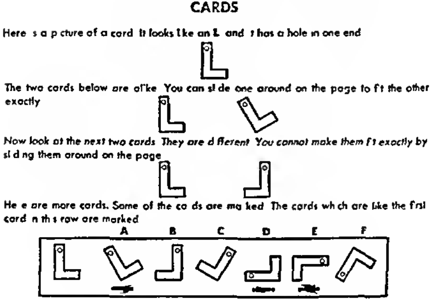



FIG 11 2 Item from Chicago Tests of Primary Mental Abilities Science Research Associates (By permission )

method, but this had not yet been done at the time the scale and its manual were published (1943) With the exception of rote memory reliabilities the coefficients reported are all between 95 and 98 (corrected by the Spearman Brown formula) These are very high The memory tests however showed coefficients of low reliability ranging from 63 to 82 Seven of these coefficients were in the 60 s two in the 70 s and three in the 80 s

Validity of the Chicago scale is reported in a manner quite different from those which are used with most other scales thus far discussed Having accepted the group-factor theory and having analyzed mental abilities into the six relatively independent factors already specified the authors proceeded on the hypothesis that the scale would be valid if the correlations between the primary abilities were relatively low, for such correlations would show relative independence of factors as required by the theory, and would thus satisfy the underlying group-factor theory of mental abilities In other words the *adopted theory* of intelligence is used as the criterion of validity rather than the more usual and generally accepted evidences of mental

development and intelligent activities. Or, to state the same thing otherwise, the criteria of the scale's validity are *internal* rather than *external*. The intercorrelations thus found range from .13 (spatial perception with rote memory) to .58 (reasoning with verbal comprehension). About three fourths of the coefficients are above .30, and about half are above .40.

**LETTER GROUPING**

Look at the groups of letters below.

AABC ACAD ACFH AACG

Three of the groups have two A's. The group which does not have two A's is marked.

Here is another problem. Three of the groups are alike in some way. Can you find three groups which are alike? Mark the one that is different.

XURM ABCD MNOP EFGH

In three of the groups the letters are arranged in alphabetical order. The first group is not in alphabetical order. You should have marked it to show that it is different.

**LETTER SERIES**

Study the series of letters below. What letter should come next?

a b a b a b a b a b c d e f

The next letter in this series should be a. The letter a has been marked in the answer row at the right.

Now study the next series of letters and decide what the next letter should be. Mark the letter in the answer row at the right.

c a d a e a f a a c d e f g

You should have marked the letter g.

FIG 11.3 *Items from Chicago Tests of Primary Mental Abilities.* Science Research Associates. (By permission.)

An individual's raw scores obtained on this scale, for each of the "primary abilities," are converted into percentile ranks which are then plotted on a profile. In the first versions of the tests, no mental age, intelligence quotient, or other *general* index was obtained, "Since the principal purpose of the present test is to obtain a profile of the six

primary mental abilities for each child . . . ."[11] It was the view of the authors of this scale that representation of mental abilities by means of a profile is not only consistent with their theory of mental organization and functioning, but that a profile is most valuable for purposes of interpretation of an individual's performance and for educational and vocational guidance.

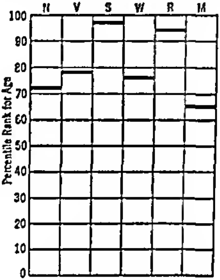



FIG. 11.4. An Individual Profile Chart—Chicago Tests of Primary Mental Abilities. Manual of Instructions, Chicago Tests of Primary Mental Abilities. Science Research Associates. (By permission.)

The data thus far available indicate, however, that taken as a whole, as a composite, the tests of "primary mental abilities" are in some instances moderately and in others poorly correlated with scholastic performance at the high-school and college levels. The data show also that the "primary mental abilities," when considered separately or in "patterns," do not differentiate significantly between persons interested in or engaged in the study of various professions (engineer-

[11] *Manual of Instructions*, p. 19. In later versions of the scales the authors provide mental age units and IQ equivalents. However, they still regard separate percentile ratings as superior to MA and IQ.

ing, science, linguistic studies, medicine, journalism, etc ) [12] Judgment must be suspended, therefore, regarding the value of these tests in specific rather than general problems in educational and vocational guidance

**KUHLMANN-ANDERSON TESTS (6TH EDITION)** [13]

**Description.** The first edition of this series of tests was published in 1927 Since that time they have undergone extensive experimentation, re standardization, and improvement so that they are at present among the superior group scales They are superior in respect to standardization procedure, both intensive and extensive analysis of results, and availability of statistical evidence presented in the manual

The entire series of scales, of which there are nine, graded according to school level, includes thirty-nine subtests Each subtest *as a whole* is placed according to its over-all relative difficulty in the age range, and the items *within* each subtest are placed in order of *their* difficulty Since intelligence levels vary considerably within any single age group, and since there *is* overlapping among different age groups, there is also overlapping (duplication) of subtests from one scale to the next Thus, the scales for the several adjacent levels include the subtests as indicated below

kindergarten, subtests 1–10
grade 1, subtests 4–13
grade 2, subtests 8–17
grade 3, subtests 12–21
grade 4, subtests 15–24
grade 5, subtests 19–28
grade 6, subtests 22–31
grades 7–8, subtests 25–34
grades 9–12, subtests 30–39

The subtest types are not unique, since they include the familiar nonverbal and verbal materials such as sequences, classification, num-

[12] See *Review of Educational Research* Vol 14, No 1, 1944, Chapter 3, Vol 17, No 1, 1947, Chapter 2, Vol 23 No 1 1953 Chapter 2

[13] By F Kuhlmann and R. G Anderson Published by Personnel Press, Inc, Baltimore, 1952

ber concepts, form perception and synthesis, word knowledge and facility, scrambled sentences, arithmetical reasoning verbal analysis, etc In the earlier levels, the materials are nonverbal, but they develop gradually into scales that are largely verbal and numerical and, finally, that are entirely so

Thus, to a considerable extent there is, from age to age, overlapping of the functions being measured by this series of subtests This is quite consistent with the stated purpose of the scale namely, to measure the levels of general mental development needed to succeed in school work In this connection it should be noted also that the authors of the scale do not recommend the use of separate subtest scores as measures of separate psychological functions or for guidance purposes But, quite appropriately, they do suggest that *significant* inconsistencies among scores on separate subtests furnish evidence of erratic performance for which causes and explanations should be sought Implicit in this position of Kuhlmann and Anderson is the acceptance of the general factor theory of intelligence This acceptance is later made explicit by them in their validation, as will be indicated below

**Validity and Reliability.** The usual criteria of validity have been used test performance of retarded, average, and accelerated pupils, intercorrelations of subtests, correlations of subtests with total scores, means, standard deviations, and ranges of IQ s at the several grade levels, correlations with quality of school work, and power to discriminate among successive age groups (age differences being fractions of a year), upon which heavy emphasis is laid

These criteria are, for the most part, reasonably well satisfied the test differences between the known groups of pupils are significant, the mean IQ s approximate what would be expected for the various age groups, the standard deviations of IQ s range between 9 5 (age 7) and 16 1 (age 14), but more than half of the S D 's are between 13 and 16 points, a table of age norms for each of the thirty nine subtests shows considerable discriminative ability between age increments of less than a year

Two thirds of the subtest intercorrelations fall in the 40 s and 50 s, while of the coefficients for each of the subtests correlated with total score, more than two-thirds fell between 50 and 81 These

coefficients indicate that there is marked communality of functions being measured by most of the subtests but that at the same time also some functions beyond the general one are being measured [14]

Reported correlations with school achievement are relatively high being from approximately 60 to about 80

Reliability was studied by means of the several familiar methods

Odd-even rel abilities (uncorrected) varied from 88 to 95

Standard errors of measurement after several retestings varied from 5 5 to 6 5 points in IQ

The test retest index of reliability is 90 for the data from which the standard errors of estimate were obtained [5]

Since there are time limits for these subtests the appropriateness of the odd-even method of determining reliability might be questioned The authors have shown however that odd-even reliability is just about as high when all time limits are removed as when the tests are timed the range of coefficients being from 80 to 95 These results indicate that the scales are essentially tests of power rather than speed

Scoring The Kuhlmann Anderson scales use the median mental age method employed with the Pintner Paterson Performance Scale (1917) The procedure is this a mental age is yielded by each of the ten subtests in a given scale the median of the ten values is taken as the over all mental age When this method is used the principle being applied is that this median value is most representative of an individual s general level of performance especially so since it is not affected by a few extremely high or extremely low subtest scores if such occur (whereas the *arithmetic mean* is affected by such ex

[14] *Apparently not much s gn ficance is attached to correlat on w th Stanford* B net results as an ind cat on of val d ty No data are presented in the manual on th s criter on although there is br ef ment on of close correspondence be tween IQ s obta ned with both scales

[15] The "index of rel ab l ty" is to be dist ngu shed from the "coefficient of rel ab l ty" The latter is derived from the correlat on of two sets of *obta ned* scores wh le the former is the correlat on between one set of obta ned scores and a set of *es mated* "true" scores for the same populat on sample The " ndex of rel abil ty" is always the higher of the two since it ind cates the *p obable h gher l m t* of correlat on rather than the correlat on that has ac tually been found for the t vo sets of obta ned scores An " ndex" of 90 (as re ported above) corresponds to a "coeff cient" of 81

tremes) From mental ages, thus found, and chronological ages, intelligence quotients are determined in the usual manner[16]

## GROUP SCALES FOR COLLEGE FRESHMEN

Some tests of intelligence have been constructed for the specific purpose of appraising abilities of individuals with special reference to the intellectual demands of college curricula, generally of the kind found in colleges of liberal arts and similar institutions, such as teachers colleges We shall briefly describe several of these, principally to familiarize the reader with the types of materials included These tests do not present any new or unusual principles in respect to construction, organization, or interpretation

American Council on Education: Psychological Examination for College Freshmen[17] This scale includes the following familiar subtests arithmetic problems, word definition, figure analogies, same-opposite (word meaning), number series, and verbal analogies These six tests are grouped into two classes (1) linguistic tests (same-opposite, word definition, and verbal analogies) which yield an L-score, (2) quantitative tests (the remaining three) which yield a Q-score

The authors state that these two separate scores may be used in educational counseling, for the linguistic tests have been found to have higher correlations with scholarship in colleges of liberal arts than do the quantitative scores This result is due to the fact that a very large portion of courses in these colleges make their principal demands upon linguistic activities and thinking On the other hand, for scientific and technical curricula, the authors hold, the quantitative tests may be more significant

The raw scores are converted into separate percentile ranks for total score, L-score, and Q-score Separate distribution tables are provided for men and women Also, it is not uncommon for a college to prepare its own frequency tables and to calculate percentile ranks of its students on the basis of their performance alone, rather than on the basis of national norms

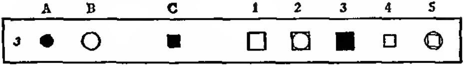
[16] The Kuhlmann Finch Tests (1952) are modeled after the Kuhlmann Anderson Tests in respect to principles and content. Published by Educational Test Bureau

[17] Developed by L. L. Thurstone and T G Thurstone Published by the Educational Testing Service annual editions A separate edition for high-school students is also available

ITEMS FROM
AMERICAN COUNCIL ON EDUCATION PSYCHOLOGICAL
EXAMINATION FOR COLLEGE FRESHMEN (By permission)

*Figure Analogies*

In Sample 3 below the rule has two parts Make Figure B of the opposite color and larger than Figure A Apply the rule to Figure C and blacken the space which corresponds to the correct answer




You should have blackened the space numbered 1 which corresponds to the large white square

*Verbal Analogies*

In each row of words the first two words form a pair The third word can be combined with another word to form a similar pair Select the word which completes the second pair On the answer sheet blacken the space which corresponds to the word you select

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| sky | blue | grass | (1) ground | (2) sod | (3) path | (4) blue | (5) green |
| ice | solid | water | (1) hard | (2) fire | (3) iron | (4) liquid | (5) boat |

*Number Series*

Find the rule in the series below and blacken the space on the answer sheet which corresponds to the next number

10 8 11 9 12 10

| 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|
| (a) | (b) | (c) | (d) | (e) |

The series above goes by alternate steps of subtracting 2 and adding 3 You should have blackened space (e) which corresponds to 13 the next number

**Ohio State University Psychological Test** [15] This scale includes three subtests same opposite (word meaning) word analogy and paragraph meaning (answering questions on each of a number of paragraphs to test the subject's comprehension and interpretation of materials read) The scale is clearly and specifically devised to test verbal intelligence

[15] Prepared under the direction of Herbert A Toops Published by Ohio College Association Committee on Intelligence Tests for Entrance Ohio State University Form 24 1950 is most recent This scale has been used with high school students also for whom separate norms have been prepared

Norms of percentile ranks are provided, based upon scores of freshmen in a large number of Ohio colleges Separate tables of norms for total scores are given for men and women Separate norms are also provided for the test of paragraph meaning The reason, presumably, for the separate norms for paragraph meaning is that thereby a subject s level of performance in respect only to word knowledge may be distinguished from his ability to interpret and think in terms of linguistic symbols

As in the case of the American Council scale, it is not an uncommon practice for colleges to prepare their own separate norms and percentile ranks when using the Ohio State scale

**College Entrance Examination Board Scholastic Aptitude Test.**[19] This scale has two sections a verbal and a quantitative (mathematical) The former section includes tests of word-opposites, word analogies, paragraph meaning and sentence completion (rather long and complex ones), the latter section includes mathematical problems, involving arithmetic, algebra, and plane geometry

The purpose of these tests is not to examine an individual s knowledge or mastery of subject matter covered in high school, even though the test items do utilize the tools of learning and thinking acquired there The verbal section is designed to measure understanding of words, skill in dealing with word and thought relationships, and ability to read with understanding and discrimination The mathematics section is designed to measure ability to handle quantitative concepts, rather than achievement in the field of mathematics On the whole, scores on the Scholastic Aptitude Test are considered to be indexes of probable success in subsequent academic work in college courses involving verbal and quantitative materials The scores on the verbal sections obviously, are found to have greater predictive value regarding performance in linguistic studies of all types and in social studies, while the mathematics section has greater predictive value in the study of physical sciences and engineering It appears that while small differences between scores on the two sections, in the case of a given person, have little significance, large differences are of im-

[19] Prepared by the staff of the College Entrance Examination Board annual editions not available for use by others. Although this is called a scholastic *aptitude* test, it has much the same content as other tests of intelligence at this level and is intended to serve the same purpose

portance in counseling students in regard to their selection of college curricula

Raw scores on this scale may be converted into percentile ranks, according to national norms Or, again, each individual institution may set up its own frequency distribution of scores and corresponding percentile ranks

ITEMS FROM COLLEGE ENTRANCE EXAMINATION BOARD SCHOLASTIC APTITUDE TEST (By permission)

*Sentence completion*

One of the most prevalent erroneous contentions is that Argentina is a country of agricultural resources, and needs only the arrival of ambitious settlers

(1) modernized (2) flourishing
(3) undeveloped (4) waning
(5) limited

Precision of wording is necessary in good writing by choosing words that exactly convey the desired meaning, one can avoid

(1) duplicity (2) incongruity
(3) complexity (4) ambiguity
(5) implications

**Yale Educational Aptitude Battery.**[20] This series of tests is made up of seven parts (1) verbal facility (the verbal section of the College Board Scholastic Aptitude Test), (2) linguistic aptitude (as measured by a test of artificial language), (3) verbal reasoning (logical inference and deductive judgment), (4) quantitative reasoning ("ability in manipulating hypothetical quantitative data so as to perceive relations or principles characterizing them and derive 'laws' analogous to, yet different from, those actually encountered in the study of the natural sciences"), (5) mathematical aptitude (mathematical problems similar to those in the College Board test), (6) spatial visualizing

[20] A B Crawford and P S Burnham *Forecasting College Achievement* (New Haven Yale University Press, 1946) The reader will note that the authors of these tests have chosen to call them an *educational aptitude* battery The contents of these scales, however are very much the same as or identical with the contents of other scales variously designated as tests of intelligence or of mental abilities or as psychological examinations Because the Yale battery is similar to these others not only in content but in purpose as well it is included here rather than in the chapter on aptitude tests

( representation of three dimensional forms by two-dimensional figures through projections and block-counting ) (7) mechanical ingenuity ( problems in gear or pulley movements structural stability and mechanical operations )

The reader is already familiar with these types of items from samples of other scales except numbers 2 4, and 7 The last of these is generally found in scales designed to test only specialized mechanical aptitudes so it will not be illustrated at this point (See Chapter 12 ) The following items illustrate tests 2 and 4 respectively

ARTIFICIAL LANGUAGE

Vocabulary

I—vlu
he it (nom )—wes
to be—jahviz
to read—skraliz
to have—dromiz
good—zeyt
book—stetsleit
word—gleit

Rules

1 Articles are not used in the artificial language
2 Verbs are not conjugated for person and number
3 Future—prefix *bli* to the verb

Sample (to be translated)

| A | B | C | Answers |
|---|---|---|---|
| I | have | a book | |

A (1) wes (2) polvlu (3) vlu (4) polwes (5) vlul   1 2 3 4 5
B (1) dromiz (2) jahviz (3) amdiz (4) somiz (5) b notiz   1 2 3 4 5
C (1) gleit (2) zepoldeit (3) zeyt (4) stetsleit (5) oveit   1 2 3 4 5

Quantitative Reasoning

| *A* | *B* |
|---|---|
| 8 | 2 |
| 32 | 4 |
| 18 | 3 |
| 200 | 10 |
| 50 | 5 |

The problem is first to discover and state the relationship between the paired numbers In this illustration it is $A = ?(B^2)$ Then when $A = 72$ what is the value of $B$?

The Yale battery differs from the other scales for college freshmen principally in the following respects it contains a more diversified and a larger number of types of items, it attempts to predict college success in more specialized areas of study (verbal, scientific, engineering), and it provides the tables, in terms of points and percentile-rank equivalents, for obtaining a profile graph for each person tested On the whole, this battery is intended to reveal individual differences in major areas and in educational promise at the higher levels which, most persons would grant, depends to a very large degree upon those mental functions that have been designated as constituting intelligence

**Evaluation of Group Scales for College Freshmen.** The reliability coefficients of the four scales discussed above are of the general order of 90 The method usually employed is the familiar one of correlating scores on odd numbered items with those on the even numbered ones

The intercorrelations presented in Table 41 are representative of the subtests used at the level of college freshmen

Since scales for college freshmen are designed for a specific purpose—selection of promising students and prediction of college achievement—their validity depends upon the success with which they perform this task The scores on these scales, therefore, have been correlated with marks obtained in college courses [22] The medians of the coefficients found in recent years are generally in the neighborhood of 50 and 60

Test scores have been used to predict survival in college, disregarding actual marks It was found that the scales are useful in helping to identify individuals at the two extremes namely, those who are most likely to complete their undergraduate work and those who are least likely to do so The test scores do not differentiate very well between individuals in the middle range of the distribution (approximately the mid-sixty percent), so far as scholastic achievement is concerned The reasons for this are that this middle group is fairly homogeneous, that numerous nonintellectual elements affect students' course marks, and that college marks themselves are neither objective nor entirely reliable The results of correlational and the 'survival' studies are not such as to warrant exclusive use of psychological tests in selection and elimination of candidates for admission to college

[22] See *Review of Educational Research* Vol 14 No 1 1944 Chapter III and the first numbers in 1950 and 1953, also Crawford and Burnham *passim*

TABLE 41

*ACE Psychological Examination*
College Edition 1948[21]
*(385 Cases at One College)*

*Intercorrelations Means and Standard Deviations of the Six Subtests Q L and Total Score*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| (1) Arithmetic | | 432 | 558 | 427 | 297 | 450 | 739 | 439 | 643 |
| (2) Figure Analogies | 432 | | 488 | 349 | 285 | 529 | 836 | 442 | 690 |
| (3) Number Series | 558 | 488 | | 323 | 283 | 498 | 847 | 421 | 682 |
| (4) Completion | 427 | 349 | 323 | | 658 | 574 | 436 | 829 | 757 |
| (5) Same-Opposite | 297 | 285 | 283 | 658 | | 563 | 351 | 904 | 766 |
| (6) Verbal Analogies | 450 | 529 | 498 | 574 | 563 | | 611 | 822 | 836 |
| (7) Q Score | 739 | 836 | 847 | 436 | 351 | 611 | | 530 | 827 |
| (8) L-Score | 439 | 442 | 421 | 829 | 904 | 822 | 530 | | 915 |
| (9) Total Score | 643 | 690 | 682 | 757 | 766 | 836 | 827 | 915 | |
| Mean | 8 16 | 15 96 | 16 58 | 16 53 | 21 16 | 26 86 | 40 70 | 64 56 | 105 25 |
| σ | 2 91 | 5 34 | 4 85 | 4 14 | 7 16 | 5 62 | 10 72 | 14 94 | 22 53 |
| *Percentile rank of means* | | | | | | | 44 | 50 | 47 |
| *Results for all colleges as reported in Norms Bulletin for 1948 College Edition* | | | | | | | | | |
| Mean | | | | | | | 41 56 | 64 38 | 105 91 |
| σ | | | | | | | 11 39 | 16 12 | 24 65 |

But they do make a valuable contribution when used in conjunction with other information and criteria concerning each candidate, such as high school marks, performance on subject matter examinations, and ratings by teachers

## ARMY GENERAL CLASSIFICATION TEST[22]

This scale, in various earlier versions, was used in the Army in World War II to classify men and women according to their abilities to learn military duties Since much of the learning in modern military service is of a technical kind emphasis was placed upon verbal comprehension quantitative reasoning and spatial perception The AGCT, therefore, adapted and utilized conceptions of testing and

[21] Communication from Educational Testing Service (By permission )

[22] Published by Science Research Associates Chicago 1947 See Staff Personnel Research Section, Classification and Replacement Branch Adjutant General s Office *Personnel Classification Tests* War Dept Technical Manual TM 12 260 rev Washington U S Govt Printing Office 1946

types of matenals that had been developed and in use for a long time prior to the war Three types of test items were employed to measure the three processes, respectively vocabulary, arithmetical problems, and block counting The items are presented in spiral form a group of vocabulary items, then a group of arithmetical problems, then blocks The sequence is repeated a number of times each group of items being of greater difficulty than the preceding one They are of the usual multiple-choice type

**Scoring.** The raw scores might range from zero to 150 These are converted into a standard score, so arranged that the mean is 100 and the standard deviation is 20 Tables are also provided for conversion of raw scores into percentile scores

TABLE 42

*Distribution of AGCT Standard Scores*[24]

(*N* = 160,000)

| Score | Percent of Sample |
|---|---|
| 41–59 | 4 |
| 60–89 | 23 |
| 90–109 | 31 |
| 110–129 | 33 |
| 130–161 | 9 |

**Validity and Reliability.** Using the odd-even method, the mean of the reliability coefficients (corrected) was approximately 95 (N varied from 639 to 3856 ) When the test retest method was used, the reported reliability was 82

The AGCT scores correlated 73 with number of school grades completed It correlated as follows with other tests Army Alpha, 90, Otis Higher Mental Ability Examination, 83, American Council on Education Psychological Examination, 79 That the AGCT scores are not in part a function of chronological age is shown by a correlation of 02 with CA (N = 4330)

Test scores were correlated with training-school marks in several hundred training programs The coefficients that follow are based upon groups of trainees who had been preselected for a particular

[24] From *Examiners Manual* AGCT, p 3 Science Research Associates (By permission )

school on the basis of education and civilian occupation The coefficients are lower therefore, than they would have been if an entirely unselected group of trainees had been used Correlations were clerical trainees 40, airplane mechanic trainees, 35, sheet metal trainees, 27, radio operator mechanic trainees 32, officer candidates (various services), 40

The ranges of scores classified according to civilian occupation of the subjects show such extensive overlapping of groups (interquartile range also 10th and 90th percentile range) that occupational differentiation would be extremely doubtful, except in instances where occupations being compared are very widely separated on the scale

The army versions of this test were constructed for the purpose of rapid and rather crude screening of huge numbers of inductees under pressure of an emergency situation The tests did this job satisfactorily But adaptation of these tests to civilian purposes is a very doubtful procedure for superior single scales and batteries of tests are available These scales and batteries which often require much time and careful interpretation are to be preferred in civilian situations that are free from pressures for speed and in which extravagance with human and material resources is discouraged

## MILLER ANALOGIES TEST[25]

Originally devised in 1926 this test has been developed to measure scholastic aptitude at the graduate-school level It consists of one hundred items with a time limit of fifty minutes but the speed factor is said to be of negligible importance The test includes analogies covering a wide variety of fields of learning and specialization Although some quantitative as well as verbal materials are used the items are predominantly verbal in character Also the relationships to be discerned within each of the items are not uniform thus the subject must reorient his analysis with each item (It is not possible to illustrate or describe the items since the test is not circulated and its use is controlled.)

Items were retained for their differentiating ability as determined by an item analysis of results obtained with 1241 college seniors

[25] By W S Miller Published by The Psychological Corporation 1947 A newer form (H) 1950 is also available The 1947 and 1950 forms are practically equivalent.

These items were then administered to 770 entering graduate students at the University of Minnesota The latest revision was made on the basis of an item analysis of their scores, the items being arranged in their order of difficulty as determined by percentage of errors These students represented a wide range of major fields of study

**Validity and Reliability.** Coefficients of reliability, found by the odd-even score method, for three groups of graduate students, were 93, 93, and 92, as corrected by the Spearman Brown formula The numbers were 100, 162, and 125

Validity is determined largely by ability of the test to predict success in graduate study The correlation coefficients between the Miller Analogies Test and numerical marks, as reported in the manual, are restricted to University of Minnesota graduate students in one field namely, education For course grades, the twenty-four coefficients ranged from 14 to 78 (With one exception, all were 35 or above ) The median was 54 Ten correlations with grades in final comprehensive examinations ranged from 28 to 54, with a median of 39 Although correlation coefficients are not reported, the data show an increase in mean test-score as honor point ratios increase Validation studies have been made subsequently at other universities The correlations found at these institutions, in other fields of graduate study, were of approximately the same magnitude

When the analogies test scores were correlated with average scores on seven parts of the Graduate Record Examination (mathematics, physics, chemistry, biology, history, government and economics, literature and fine arts) the coefficients were much higher These ranged from 64 to 84, with a median of 775 Correlations with the advanced Graduate Record Examination (fields of specialization) yielded coefficients of 81 and 79 for major students, respectively, in chemistry and languages and literature Correlations with the verbal parts of the Graduate Record Examination were from 74 to 81, the median being 80

This array of coefficients adds up to the conclusion that the Miller Analogies Test is a useful additional source of evidence in regard to a person's ability to pursue graduate study Combined with results of the Graduate Record Examinations, the value of each is increased The lower coefficients obtained with actual course grades are very

probably due to the variations in marking, the differences in difficulty of courses, specialization of interests and abilities and influence of nonintellective factors that affect a student's level of performance

This analogies test is a carefully constructed instrument, devised for use with a difficult problem namely differentiation among individuals in a rather highly selected group The ability tested by means of these analogies is one factor in the prediction of success in graduate studies namely, ability to learn verbal and other abstract concepts and course materials What the analogies do not evaluate (nor do they purport to do so) is original thinking and constructive research ability—both of which should rank very high in graduate studies

## OTHER GROUP SCALES

It is our purpose in this section, to refer briefly to several scales which will help to give the reader a more nearly adequate conception of representative group instruments

**Institute of Educational Research Intelligence Scale CAVD** This test was developed at Columbia University under the direction of E L Thorndike and frequently is designated by his name [24] The scale gets its name (CAVD) from the fact that it proposes to measure intellect specifically by means of four kinds of mental tasks completion (C), ability to supply words so as to make a statement true and sensible', arithmetical problems (A), vocabulary (V), ability to understand single words, directions (D), ability to understand connected discourse as in oral directions or paragraph reading

The distinguishing feature of this scale is that the items are arranged in order of difficulty, providing seventeen different levels in each of which the tasks of any one subtest (e g., C) are of nearly equal difficulty Also, the steps between levels are of approximately equal difficulty The lowest level is suitable for three-year-old children while the highest levels are intended for superior adults Thus the scale is designed to test the same functions in a continuum from early childhood through adulthood

The range of difficulty in the same category of mental tasks may be illustrated by the following examples

[24] Published by Bureau of Publications Teachers College Columbia University, 1925 Norms are not available for levels below ninth grade population

*Completion, lowest level*
You are sitting on a
*Completion, one of the highest levels*
Throughout the river plains of northern India, two harvests, and, some provinces, are each

*Arithmetic lowest level*
Counts two pennies
*Arithmetic one of the highest levels*
A factory earns $70 a day for its owner when it is working full capacity and $15 a day when it is working half capacity In how many days will it earn $1,000 if two days out of every three are only half capacity?
*Vocabulary, lowest level*
Show me the horse (to be indicated in a series of pictures)
*Vocabulary one of the highest levels*
Accolade [means] (1) salutation, (2) anchovy, (3) procession, (4) bivouac, (5) acolyte
*Directions lowest level*
Make a ring like this,' showing act
*Directions one of highest levels*
A rather long paragraph, entitled "The American State," is read The following is one of the questions asked "To what may we attribute the similarity between the plans of certain cities and the arrangement of the States?'

Results obtained with the CAVD scale have been correlated with those obtained by the same persons on other group scales and on the Stanford Binet The coefficients obtained were high, and in some instances very high However, the time required to take the CAVD scale is much longer than that required for the others with which it correlates so highly Due to the time factor, this scale is not so widely employed as some of the others

**The Thorndike Intelligence Examination for High-School Graduates.** This test is extremely long (requiring about three hours) and laborious to score, as well as expensive[27] It includes tests of highly specialized information, arithmetic, algebra, and paragraph meaning (which constitutes more than two thirds of the examination) The information and algebra tests may be criticized as being dependent upon highly specific school learning, and the tests of paragraph inter-

[27] By E L. Thorndike Published by Bureau of Publications, Teachers College, Columbia University, most recent series, 1931-1936

pretation may be criticized as being dependent upon a high level vocabulary which requires unusual educational opportunities, either formal or informal Thorndike, however, intends that the examination shall be used with "candidates who have had good educational advantages and who know English as a mother tongue" The presumption, then, is that, having had these advantages, prospective college students will manifest intellectual competence and promise in terms of their abilities to deal with such mental tasks as are presented in this examination

**Henmon-Nelson Tests of Mental Ability.** Though one of the older instruments, this test is still rather widely used[*] Standardized at three educational and age levels (grades 3–8, 7–12, and 12–16), the items are arranged in "spiral omnibus" form These scales, intended to measure general scholastic intelligence, are weighted with verbal materials Of secondary importance (in terms of numbers) are the quantitative items, and of distinctly minor significance are the nonverbal (spatial relations) items, there being, for example, only 10 of these in 90 at the level of grades 7–12 The verbal and numerical items are of the familiar variety synonyms, definitions, analogies scrambled sentences, number sequences, arithmetical problems, etc

The Henmon-Nelson scales rank among the superior group tests of general ability, when used with subjects who are not under language or educational handicaps Although these are scales upon which a great deal of careful work has been done, they are now in need of revision in respect to norms and content

The following four items illustrate how a "spiral omnibus" test is arranged

1 Which word does not belong with the others?
 (1) Ida, (2) Paul, (3) Lucy, (4) Janet, (5) Edith
2. Better is to *good* as *worse* is to
 (1) very good, (2) medium (3) bad (4) much worse, (5) best
3 1, 6, 11, 16 . . . . , . . . . , 31 What two numbers should be on the dotted line?
 (1) 21 and 26 (2) 17 and 25, (3) 26 and 29, (4) 22 and 27, (5) 20 and 25
4 *It was raining too hard to ....... out A word for the blank is*
 (1) comment (2) gather, (3) venture, (4) summon (5) render

[*] By V A C Henmon and M J Nelson Published by Houghton Mifflin, 1931–1950

**The Pintner Tests.** Another series is that of which Pintner was the senior author[29] The *Pintner-Cunningham Primary Test* (1946) covers the range from kindergarten through the first half of grade 2 The *Pintner-Durost Elementary Test* (1940) is devised for the last half of grade 2 through the first half of grade 4 The *Pintner Intermediate Test* (1938) is for the latter half of grade 4 through the first half of grade 9 The *Pintner Advanced Test* (1938–1939) begins with the ninth grade and continues through adult levels

**The Otis Group Intelligence Scale (1919).** This test provides two examinations[30] The Primary Examination is designed for the range extending from kindergarten through grade 4 The Advanced Examination extends from the level of grade 5 through grade 12 Otis has also devised several other forms of his tests, which differ from the foregoing chiefly in respect to their mechanical features, some are designed to be "self administering," and others to facilitate "quick scoring"

There are other group scales which are as well constructed, as valuable, and as widely used as some described in this chapter It has not been our purpose to present a complete listing and description of group scales, it is unnecessary to do so, since those that have been described herein contain all the essential features of current group tests It has been our purpose, rather, to familiarize the reader with the organization and content of group scales, their statistical merit, their scoring techniques, and with the major similarities and differences that exist among them

## EVALUATION OF GROUP SCALES

**Comparison with Individual Scales.** Group scales were developed to permit the testing of large numbers of persons at one time On the whole, therefore, they are not so useful as are individual scales (e g , Stanford-Binet and Bellevue) in studying an individual case For when a group scale is used, it is not possible to observe a person's approach to the solution of problems, nor his behavior under success and failure Nor is it possible to evaluate the *qualitative* characteristics of his responses, since group scales are scored quite rigidly Further-

[29] Published by World Book Co
[30] Published by World Book Co

more, it is difficult—in fact, practically impossible—to know whether an individual is exerting his maximum effort when taking a group examination Thus when a group scale is given, it is possible to report the test results only in terms of numerical indexes (plus profiles, at times), whereas during an individual examination the psychologist is able to make behavioral and qualitative observations of considerable value

Practically all group scales, below college level, have been validated against individual scales—especially the Stanford Binet—as one of the principal criteria This fact in itself is a recognition of the merit of the individual scale, the quality of which the group scale is trying to approach as closely as possible Other criteria of validity are the familiar ones discussed in earlier chapters

In discussing the definitions and analyses of intelligence, we stated that one deficiency of all tests is that they do not measure the creative aspects of intelligence, nor do they directly measure the insights that come from experience ("wisdom," "judgment"), or productive thinking, or the intellectual originality of an individual This deficiency is more marked in group than in individual scales because of the rigidity of scoring the former

**Theoretical and Statistical Bases.** Most group scales are based, implicitly at least, upon the "general factor" theory of intelligence, for most of them undertake to sample a person's mental activities by *means of several kinds of tasks and* then *to rate the individual* by means of a single index A few scales are based upon the group-factor theory

Since many group tests, of varying quality, have been published, it is essential that prospective users examine the manuals closely to determine which of these satisfy the standards that should be demanded of them The reader is already familiar with the standards and methods of establishing reliability and validity These should be rigorously applied to group tests In this connection it is essential that the manual state which method was used in determining reliability, especially if the speed factor seems to be a significant one

*Since group tests for children and adolescents are used* primarily to assist in dealing with educational problems, it is essential that the scale s predictive efficiency, with regard to school work and progress, be reported as one criterion of validity

Scores from a scale as a whole are more reliable and more valid than subtest scores A distinction should be made, therefore, between subtest reliability and validity, on the one hand, and total scale reliability and validity, on the other This distinction is especially pertinent when a scale's subtest scores are to be used for differentiating and diagnostic purposes

The manual should give not only the size of the standardization population sample, but the characteristics of that sample should be specified namely, geographic and socio-economic distributions, range of ages, range of ability levels, range of school levels, and sex distribution

**Criteria of Evaluation.** In evaluating a group scale with a view to its possible usefulness in a given practical situation or in the solution of a theoretical problem, it is customary to use the following criteria

It must be sufficiently *valid* and *reliable*

The range of *norms* must be adequate for the group for which the scale is devised

The *item difficulty* in each subtest must be of sufficient range to differentiate between the various levels of ability Individuals at the lowest and highest levels should be able to obtain representative scores

In general, the *range of ability* to be tested (ages and school grades) should be restricted rather than all inclusive By restricting the range, a given number of items and a given length of time can be used for a more thorough and accurate examination than if a scale of the same length were employed to cover a wider range In the latter instance, the test items would have to be spread more thinly

*Length of the scale* must be adequate In time required, scales vary from about one half hour to three hours, depending upon levels for which they are intended The great majority of scales require one and one half hours or less Increase in length, to an optimal point, adds to the validity and reliability of a scale, for errors of measurement are decreased (better sampling) as length is increased to an optimal point Judging from current practices, based upon experiment, optimal lengths appear to be about a half hour at the level of kindergarten and primary grades, about forty five minutes at the level of elementary grades, and up to about an hour or an hour and a half at higher levels

*Simplicity of responses* is frequently regarded as an asset in group tests For some purposes—when group trends are sought, rather than individual performance—this is an asset simply because scoring is facilitated But, as already pointed out, such simplicity and conse-

quent rigidity may limit the value of tests when evaluation of an individual s responses is desired

*Simplicity of scoring* is also frequently considered to be an asset since it is actually a result of simplicity of responses The same com ments apply here as above

*Ease of administering* a group scale is desirable Frequently group scales have to be given by relatively inexperienced persons it should therefore be possible to train them in a brief time to administer the scale accurately and with precision Also simplicity of instructions and procedures in giving an examination to a group reduces the pos sibility of confusion and misunderstanding on the part of individuals in the group

The *examiner s manual* should be clear and complete in respect to standard zation procedures and results nature of the content direc tions for administering and scoring norms and interpretation of results

The *content* of the tests should be *interesting* to the groups for whom the scale is intended

The *content* of the tests should be *appropriate* to the subjects being examined That is to say the psychologist must determine whether or not in a given instance it is desirable to use a scale which is entirely verbal or entirely nonverbal or verbal and quanti tative or mixed His choice of scale will depend upon who are the subjects to be tested and the purpose for which the test is being given

## USES OF GROUP SCALES

Without going into details we wish at this point only to men tion the uses to which group scales have been put

In schools they have been used for purposes of general survey, ability classification of pupils and guidance Under general survey, studies have been made of the following range and distribution of mental ability age and grade overlapping of ability, differences be tween pupils in various schools within the same community differ ences between pupils in different school systems differences between pupils in the several high-school curricula the effects of different methods of instruction upon pupils at the several levels of ability relations between intelligence test ratings and school achievement in general and in specific school subjects and comparisons of city, town and country children

In classifying pupils according to ability level for the purpose of differentiated instruction a test of mental ability is of course basic though it should not be the only criterion

Since relatively very few schools include a qualified psychological examiner on their staffs, and since extensive individual examination is costly and time-consuming, group scales are being used for most guidance purposes However, in view of the fact that group test ratings may indicate only the approximate level of an individual's mental ability they must be used in conjunction with other available evidence obtained from school records, teachers' reports, objective achievement tests, and interviews But there is no doubt that psychological test ratings, correctly obtained and interpreted, tell us much more about a pupil s mental alertness and organization of abilities than could be ascertained without their use

Group tests have been applied extensively to a large number of theoretical and practical problems of psychological educational, and sociological significance such as individual differences in relation to sex, racial, and national membership, mental levels and characteristics of special groups, such as the mentally deficient, the gifted, and the delinquent, employee selection for jobs requiring different levels of ability, family similarities and the inheritance of intelligence, effects of changed environment upon mental level, the nature and course of mental development, the nature and organization of intelligence, constancy of the IQ and prediction of ability, and problems of theory and technique, such as the relationship between "speed ' and "power as aspects of intelligence Then, of course, there was the vast use of group scales in the armed forces, during World War II, for "screening and classification of enlisted and commissioned personnel

The foregoing enumeration is not exhaustive, but it suffices to show the wide range of application of group tests of mental ability, and it explains why tests are under continual scrutiny in an effort to increase their validity and reliability

# 12.

## APTITUDE TESTS: MECHANICAL AND CLERICAL

### DEFINITION AND EXPLANATION

An *aptitude* is a condition or combination of characteristics indicative of an individual's ability to acquire with training some specific knowledge, skill, or set of responses, such as the ability to speak a language, to become a musician, to do mechanical work, etc. An aptitude test, therefore, is a device designed to indicate a person's *potential ability for performance of a certain type of activity of a specialized kind and within a restricted range*.

Aptitude tests are to be distinguished from those of general ability, such as we presented in earlier chapters, and also from tests of skill or proficiency acquired *after* training or experience. They should be distinguished, too, from educational achievement tests which are designed to measure an individual's quantity and quality of learning in a specified subject of study after a period of instruction.

The reader should note that aptitude is differentiated from skill and proficiency. *Skill* means the ability to perform a given act with ease and precision. *Proficiency* has much the same meaning except that it is more comprehensive, for it includes not only skills in certain types of motor and manual activities, but also in other types of activities as shown by the extent of one's competence in language, bookkeeping, history, economics, mathematics, human relations, etc. We may speak of one's degree of proficiency in any type of performance. On the other hand, when we speak of an individual's aptitude

for a given type of activity, we mean the capacity to *acquire* proficiency under appropriate conditions, that is, his potentialities at present, as revealed by his performance on selected tests which have predictive value

Furthermore, when we speak of a person's aptitude for a specified activity, we do not make any assumptions regarding the degree to which they depend upon innateness or acquisition In giving an aptitude test to a person we desire to obtain a measure of his promise or essential teachability *in a given area* While we make no assumptions regarding the roles of nature and nurture in this matter, we, as clinicians or guidance counselors, cannot ignore that person's past experience in evaluating his performance on aptitude tests For example, one method of measuring mechanical aptitude is by means of a mechanical assembly test, utilizing various common objects such as a bicycle bell and a door lock It is inconceivable that a boy who in the past has had opportunity to manipulate such objects will not achieve a higher score than if he had not had such experience Testing instruments measuring engineering aptitude include, for example, tests of simple mathematical relationships, scientific vocabulary, common scientific principles, and problems of practical mechanical insight Here again, an individual's performance will be influenced by his previous experience This aspect of aptitude testing and interpretation will become clearer as the reader becomes acquainted with the nature and content of aptitude tests

The principles underlying aptitude tests are the same as those employed with tests of intelligence in respect to sampling of performance, population samples, and standardization techniques (including reliability and validity) Therefore, we shall not present the several aptitude tests in statistical detail It will be our purpose, rather, to describe the kinds of activities or functions most commonly examined by available tests of this type

## TESTS OF VISION AND HEARING

Quite aside from the general desirability of good vision and hearing, there are numerous occupations and forms of learning in which one or both are essential at a high level, thus, they are aspects of certain aptitudes Sensory deficiencies, furthermore, may adversely affect an individual's achievements in schoolwork or in his social and emotional adjustment Hence, in some cases they might play a sig-

nificant part in clinical work and in vocational and educational guidance Tests are available for visual acuity, color vision, and auditory acuity

**Color Vision Tests.** All such tests depend upon the principle that color-deficients confuse certain groups of hues, *inter se*, while a normal person distinguishes them Thus, one set of charts is so devised that persons with unimpaired color vision should see certain bars, or arms, radiating from the centers of the circles In one of the circles, for example, a person having unimpaired color vision will see two radiating arms one green and one red A red-blind eye will see only the green, a green blind eye only the red, and the red-green blind eye will see neither Another set of charts is so devised that one with normal color vision will see certain numerals, whereas the color-deficient will not The weakness of these tests lies in their requirement of a standard illuminant for testing—a condition which is almost never in effect For example, in the Navy, during World War II, about fifty percent of all color-deficients remained undetected after one to five medical examinations, and despite instructions to examiners to exercise great care as to illumination However, new research is now complete on a color vision test which remains diagnostically stable despite variations in illumination [1]

**Tests of Visual Acuity.** These have been taken by nearly everyone, the most familiar being the crude Snellen Chart On this chart are printed rows of letters, varying in size, to be read by the subject Each row and size has been standardized as recognizable at a specified distance by the "normal eye " Visual acuity is expressed as a fraction the numerator is the distance the subject stands from the chart (usually 20 feet), and the denominator is the "distance value" of the smallest letter that can be read by the person being tested ' Distance value ' of a given size is the distance at which a letter of that size can be read by the ' normal eye " Thus, if the smallest letter read by a person standing at a distance of twenty feet is read by the ' normal eye ' at forty feet, that person's vision (in that eye) is given as 20/40 The present Snellen Chart, though still used is a very inadequate

[1] E Freeman "An Illuminant Stable Color Vision Test I *Journal of the Optical Society of America* Vol 38 1948 pp 532–538 This test is distributed by The Psychological Corporation

test, for it will detect only the myopes, but not the hyperopes and presbyopes, nor those heavily handicapped by muscle imbalance

For some years, the only available device for more thorough nonclinical testing of near and distance visual functions was the Keystone "Telebinocular," the accuracy and dependability of which have been seriously questioned Just recently, however, a new, much more efficient device, known as the Protometer, has been developed It is administerable by nonclinical persons for testing both near and distance visual functions (acuity, muscle balance, depth perception, and color) [2]

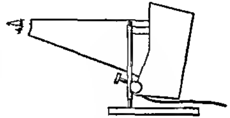


FIG 12 1 *The Protometer, designed and developed by Ellis Freeman (Patented ) (By permission )*

*The Protometer is designed for rapid and comprehensive testing of vision where numerous individuals are involved, as in schools, colleges, industries, and the armed forces The Protometer gives, among other data, monocular acuity, binocular acuity, and muscle balance, both for distance and for near—all under proper conditions of illumination and viewing maintained at a constant level The Protometer discloses cases of serious impairment of vision where the trouble is not with acuity but with the failure of the two eyes to work in coordination*

*The speed of operation of the Protometer is due to the fact that it is basically two Brewster stereoscopes the optical system of the one for distance over the optical system of the one for near As long as the subject is taking the distance test, it alone is illuminated When the target for the near test is to be viewed, the target light for the distance test is automatically extinguished while the target light for the near test is automatically turned on The tester operates only one control knob, which presents in sequence the test series for both distance and near and at the same time controls the illumination The second knob is merely for adjusting the eye height of the ocular to the subject, when this is occasionally necessary, the entire instrument is raised or lowered on a rack or pinion*

*The Protometer, weighing 10 pounds, is easily portable and can be used wherever it may be plugged into current The operator is a teacher, nurse, clerk, or other nonprofessional person who is taught to operate the instrument read the questions from the record card, and enter responses The record card itself informs the operator of the results, which may indicate*

[2] Distributed by Freeman Technical Associates 1206 Benj Franklin Dr, Sarasota, Fla.

*either satisfactory vision or deficient vision requiring referral for professional examination The* **Protometer** *test is administered within two minutes per subject*

*The test material of the* **Protometer** *consists of the basic tests used in professional refraction and has the same high precision For this reason it is extremely dependable as a preliminary diagnostic device It is not equipped with means for complete diagnosis and prescription of correction For these latter a full scale professional referral is necessary*

Since examinations of men in the last war have shown that almost ten percent of all males are color-deficient in some degree it seems desirable to test all school children very early for this function Deficients could for example be diverted from trying to become artists geologists clothes designers etc It is wise also to use a dependable color test in personnel divisions of department stores and of some industries in the former to avoid placing color-deficient sales persons in the wrong departments in the latter to avoid placing a colorblind individual on say radio or other wiring which requires discrimination of color code For a job in which some color deficiency can be tolerated, the demands of the job in regard to color discrimination should be determined as well as the candidate s degree of deficiency

Good vision as such is desirable without question Hence in schools and industry there should be screening by means of a reliable device to find those who need correction or whose color deficiencies need to be given consideration in education

*Auditory Acuity* This sensory function is commonly measured (1) by means of whispered words testing the hearing of consonants and vowels such as the Andrews Whispered Speech Test or (2) with an audiometer The former consists of one hundred numbers which are whispered at a specified distance A person s score is the percentage of numbers correctly heard divided by the normal percentage The audiometer provides stimuli which are superior to whispered speech for they are of measured intensity and unaffected by acoustics of the room in which the test is being given The device consists of a disc and phonograph with magnetic reproducer that presents numbers of two or three digits each The subjects hear the numbers through head phones and write them down each ear being tested separately Successive numbers are reproduced at decreased intensity at small uniform steps until a minimum intensity scarcely audible to the "normal ear is reached The cycle from maximum to minimum intensity is repeated four times for each ear Of the eight series of numbers four

are spoken in a male voice and four in a female's There are tests, also, of sentences, words, and pure tones

It is obvious that these tests of vision and hearing do not measure a person's aptitude for specific types of learning and activity For certain kinds of learning and activity, however, a given degree of visual or auditory acuity is essential In that sense, then, these devices may constitute a part of a battery of tests which, taken together, are used to measure a particular aptitude

## MOTOR AND MANUAL TESTS

**Tests of Strength of Grip.** One of the oldest instruments for the measurement of individual differences in the psychological laboratory is the hand dynamometer for measuring *strength of grip* The instrument consists of an inner and an outer handle, a dial, and a pointer The subject grips these handles so that the second phalanges of the fingers press against the inner handle, while the outer handle presses against the heel of the hand The subject then squeezes as hard as possible Strength of grip is measured in kilograms After many experiments, it appears that in psychological work this instrument is useful principally as one device for determining degree of handedness and rate of fatigue Since these two traits are involved in certain activities and occupations, they are relevant in some aspects of aptitude testing

**Tests of Reaction Time.** Reaction time is the time interval between the onset of a stimulus and the beginning of the person's overt intentional response The particular stimulus and response to it are prearranged in an experimental situation For example, the subject may be instructed to tap a telegraph key immediately upon perceiving a red light, the elapsed time between stimulus and response being electrically recorded in terms of thousandths of a second It is possible to devise a variety of tests, their particular character depending upon which sensory and motor functions are to be measured This type of test obviously is intended to measure speed of response in situations demanding immediate reaction, as in certain machine operations and in driving an automobile

**Tests of Manual Dexterity.** In order to achieve competence in activities requiring manual dexterity, speed of gross movements of hand and arm, manual rhythm and coordination, and finger control and

coordination are necessary in varying degrees For each of these purposes several tests have been devised which vary in detail but are fundamentally alike Gross movements of hand and arm may be measured in terms of speed with which the subject picks up and places cylindrical blocks in holes in a board Finger dexterity and coordination, necessary in rapid and accurate manipulation of objects, may be tested by measuring the rate at which an individual, with fingers or tweezers, is able to pick up small cylindrical metal pins or wooden pegs, of different shapes, and place them in the holes of a tray (See Fig 12 2 ) Hand precision is measured by the accuracy with which

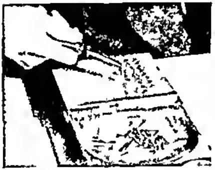


FIG 12 2 The plier dexterity test shown here is useful in evaluating skill in the use of small tools and in general in evaluating aptitudes involving finger dexterity The tray contains metal pegs which must be placed in the small holes in a prescribed order The score is based upon time required to complete the task Sometimes the time required to remove the pegs is also included in the score (Acme Photo )

a metal stylus can be placed into holes of small diameter cut in metal and electrically connected Contacts of the stylus with rims of holes are electrically recorded and constitute the measure of inaccuracy Occasionally, also a paper-and pencil test includes tasks designed to measure hand precision, such as speed and accuracy of tracing a path speed of tapping, and placing a prescribed number of dots within a small circle [3]

Other tests of manual dexterity follow the same general form but some are more complex For example, the *Small Parts Dexterity Test* [4] consists of a metal plate having two sets of 36 holes each (six rows and six columns) One set of holes is threaded, while the other is smooth The testee uses forceps to place a pin in each smooth hole, then a collar over the pin In the threaded holes the testee places a small screw then tightens it down with a screwdriver The stated

[3] For example the MacQuarrie Test for Mechanical Ability Published by California Test Bureau 1943

[4] By J E. and D M Crawford The Psychological Corporation 1949

purpose of this test is to measure a combination of perception and dexterity, in terms of rate of performance

*The Stromberg Dexterity Test*[5] also is a device the purpose of which is more complex than the simple measurement of manual dexterity It consists of a tri-colored formboard (6 rows and 9 columns), into which flat, cylindrical disks, variously colored, are to be placed by the subject in a prescribed order It appears that this test involves not only manual dexterity, but also gross color perception and a rather elementary level of nonverbal classification

Since manual dexterity scores on tests of this type are affected, in varying degrees, by the subject's lateral dominance (the preferred use and superior performance of one side of the body or the other), it is often desirable to use tests of band and eye dominance This procedure is particularly indicated if we are concerned primarily with analyzing and understanding the person being tested, rather than with making selections from among candidates for a particular job[6]

Coordination and rhythm of hand movements have been tested by means of a card sorting test in which the subject uses one hand at a time or both hands together in dropping playing cards through slots A more recent device is the two hand coordination test in which the individual attempts to move both handles of a mechanism simultaneously in such a way as to keep an upper disk over the lower one, which moves in an unpredictable manner[7] Another two handle device is employed in testing a subject's ability to follow an irregular path without touching the sides[8]

During World War II, numerous psychomotor tests were used by army and navy psychologists to assist in the selection of men for specific types of training, especially in the air forces[9] These tests involved more difficult operations than those described above, often requiring rapid and complex sensory-motor coordination, such as the

[5] By E L Stromberg The Psychological Corporation 1951

[6] See for example, A J Harris Tests of Lateral Dominance, W R Miles *The A B C Vision Test. The Psychological Corporation*

[7] A W Melton "The Selection of Pilots by Means of Psychomotor Tests," *Journal of Aviation Psychology* Vol 15, 1944 pp 116–123

[8] For detailed summary see G K Bennett and R M Cruikshank *A Summary of Manual and Mechanical Ability Tests* New York The Psychological Corporation 1942

[9] Staffs Psychological Research Unit No 2, and Department of Psychology, School of Aviation Medicine "Research Program on Psychomotor Tests in the Army Air Forces" *Psychological Bulletin* Vol 41, 1944 pp 307–321

following the use of both hands simultaneously in manipulating two lathe type handles to follow a target which moves in an irregular path obtaining patterns of lights by manipulating stick and rudder in a simulated airplane cockpit, reacting to four different relative positions of a red light and a green light by pushing one of four switches arranged in a square pattern before the subject, moving a wheel resembling an airplane control in and out of its shaft in order to hold a horizontal bar in the center of a circular aperture causing a beam of light to follow a given course when the horizontal movement is controlled by one lever and the vertical by another

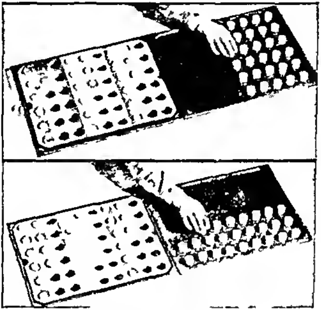


FIG 12 3 *The Stromberg Dexterity Test* The Psychological Corporation

The tests of sensory capacity and those of motor and manual dexterity developed prior to and during World War II have been moderately useful in connection with selecting persons for specific types of training or for particular jobs The functions and activities measured by these sensory and motor tests are it appears, practically unrelated to the mental functions measured by tests of general ability, for the many correlational studies made between sensory and motor tests, on the one hand and tests of intelligence, on the other, have

yielded coefficients which are very low, some being so low as to be negligible In fact, on occasion a low negative coefficient has been found It has been concluded, therefore, that these two types of psychological instruments measure functions which are largely independent of each other

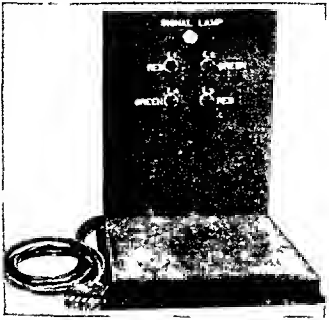



FIG 12 4 *The Discrimination Reaction Time Test This test was designed to measure how quickly individuals make differential manual responses to visual stimulus patterns differing from one another with respect to the spatial arrangement of their component parts The test requires that the candidate react by pushing one of four toggle switches in response to the lighting of a red and a green signal lamp The position of the red lamp with respect to the green determines which of the four switches should be pushed*

*A front view of a single test unit with designations of lights and switches is shown in the figure The four stimulus lamps two red and two green (L1 L2, L3 L4), are arranged in the form of a square on the vertical panel facing the candidate The stimulus to which the candidate must react by operating one of the four toggle switches is the simultaneous lighting of one of the red lights and one of the green lights*

*If he operates the correct switch, the white signal lamp (which lights on every trial) is extinguished immediately, signaling the candidate that he has made the correct response The colored lights do not go out until they have been on for 3 seconds, regardless of how quickly the correct switch has been pushed*

*The four spring return toggle switches (S1, S2, S3, S4) are so set that the candidate must push each one in a different direction The four directions of movement correspond to the four signal patterns formed by the lighting of the red and green lamps Thus, if L1 and L4 are lighted, the red is 'up' with respect to green, and the upper switch S1, must be pushed up If L3 and L4 are lighted, the red is to the right of the green, so the switch on the right, S2, must be pushed to the right The time taken to operate the correct switch on each of a series of test trials is accumulated on an electric stop-clock and constitutes the candidate's score*

(From Apparatus Tests, Report No 4, Army Air Forces Aviation Psychology Program, edited by A W Melton U S Government Printing Office, 1947 )

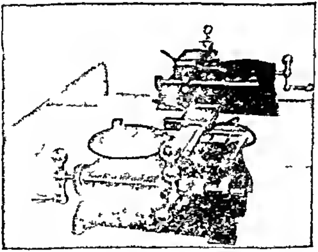


FIG 12 5 *The Two-Hand Coordination Test*

*The Two-Hand Coordination Test was designed to measure a candidate's ability to coordinate the movement of both hands He is required to control*

*the movements of a target follower in response to a visually perceived target moving at varying rates along an irregular pathway*

*A single test unit as seen from the candidate s position is shown in the figure Two handles which he manipulates are seen in the foreground and at the left Rotation of the upper handle causes a contact point which is mounted on the leaf of a micro-switch to move toward the candidate with counter clockwise rotation and away from the candidate with clockwise rotation Rotation of the lower handle in a counterclockwise and clockwise direction causes the contact point to move to the left and right respectively Rotation of both handles simultaneously causes the contact point to move in any desired direction in the plane of movement of the target A candidate s task is to manipulate the controls in such a way as to keep the targetfollower on top of a round brass button (the target) as it moves along an irregular clockwise path When the contact point is on the target button the microswitch is closed and current flows to an electric clock located on a remote control desk The time which is accumulated on the clock during a series of eight 1 minute trials indicates the efficiency of the candidate s performance*

(From Apparatus Tests Report No 4 Army Air Forces Aviation Psychology Program edited by A W Melton U S Government Printing Office 1947 )

## TESTS OF MECHANICAL APTITUDE

The capacity designated by the term mechanical aptitude is not a single unitary function It is a combination of sensory and motor capacities such as those already briefly described plus perception of spatial relations the capacity to acquire information about mechanical matters and the capacity to comprehend mechanical relationships Thus tests of mechanical aptitude are designed to measure capacity and performance on a higher level of organization than are those of sensory motor capacity and dexterity

*The Assembly Test of General Mechanical Ability*[10] devised by J L Stenquist (1923) the first of its kind and now of little more than historical interest was intended to measure a person s ability

[10] C H Stoelting Co Chicago

FIG 12 6 *Bi Manual Planned Pursuit Test*

*The Bi Manual Planned Pursuit Test was designed and developed to measure ability to coordinate the activities of both hands by a systematic shifting of attention The test consists of an irregular polished brass pathway which moves beneath two pointers The pointers are separated by a distance of 8 inches The pointers are adjustable by the candidate by means of two vertical handles the candidate being required to keep the pointers (one with each hand) in contact with the moving pathway In view of the fact that a limited amount of the pathway is visible prior to reaching the contact pointers it was believed that a certain amount of planning could occur and would in part determine the score on the test The test consists of six 1 5 minute trials Rest periods of unspecified duration probably about 30 seconds occurred between trials The score is the length of time during which both pointers are on the pathways*

(From *Apparatus Tests* Report No 4 Army Air Forces Aviation Psychology Program edited by A W Melton U S Government Printing Office, 1947 )

to put together the parts of mechanical devices among them a bicycle bell, a double action hinge, a door lock, and a mousetrap This test consisting of three series, is constructed for use with individuals covering the age range from children in the lower grades through adulthood

FIG 12 7 Triform Pegboard Test

Each test apparatus consists of two pegboards 11 25 inches wide and 28 inches long each containing 48 holes in 12 columns of 4 holes each with a depth of 11/16 inch Figure 12 7 is a photo graph of one of these peg boards There are 16 holes of each of three shapes round square, and triangular Corresponding to these holes are 16 pegs of each shape The 16 round pegs are painted red and have a diameter of 9/16 inch the 16 square pegs are painted yellow and have sides 7/16 inch in length the 16 triangular pegs are painted blue and have 3/4 inch sides All pegs have the same length which is 1 and 17/16 inches In one of the peg boards the various shaped holes are scattered randomly throughout the board while on the other board they are grouped in three sections the round holes on the left third of the board which is painted red the square holes in the center section which is painted yellow and the triangular holes in the right section which is painted blue

The boards are placed in front of candidate The board in which the holes are in irregular order contains the pegs and is placed above the board in which the holes are grouped by shape The task of the candidate is to transfer the pegs in a standard order from upper board to lower board

The test is designed for administration in a 15 minute test period one minute of which is required for preliminary instructions There are six 40 second trials with 60 second intervals between trials for replacing the pegs in their original positions and for rest The first three trials are performed with the right hand and the last three trials with the left hand There is no practice trial Scores are recorded for each trial The score is the number of pegs placed in the bottom board during the 40 second period No credit is given for a peg in the candidate s hand at the signal Stop Credit is given for all pegs placed even though in improper order

(From Apparatus Tests Report No 4 Army Air Forces Aviation Psychology Program edited by A W Melton U S Government Printing Office 1947 )

The Stenquist tests have been revised and extended at the University of Minnesota (1930) and are known as the *Minnesota Me-*

*chanical Assembly Test.*[11] In principle, these are essentially the same as Stenquist's tests, some of the same mechanical devices having been retained, with new ones added. Performance on these tests—scored in terms of rate and accuracy of work—has been found useful in predicting success of junior high-school boys in shop courses. Also, facility in these assembly tests has been found by some investigators to be one significant indication of a person's aptitude for a number of occupations such as machinist and auto mechanic.

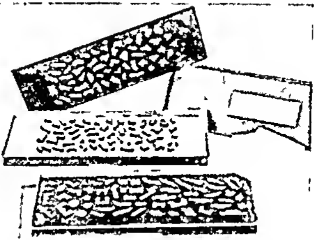


FIG. 12.5. *Minnesota Spatial Relations Tests. The upper and lower parts represent the formboards which are filled with the pieces represented in the middle part.* (Educational Test Bureau. (By permission.)

*The Minnesota Spatial Relations Test* (1930) consists of a series of four boards, each of which has 58 cutouts of various shapes, many of them unusual.[12] The subject's task is to replace these in their correct holes in the board. Evidence indicates that persons engaged in mechanical occupations tend, as a group, to earn higher scores than do persons in nonmechanical occupations. This fact, it appears, is a principal justification for use of the test as a measure of mechanical aptitude. Some critics of the test have concluded that it is adequate

[11] D. G. Paterson, et al., *Minnesota Mechanical Ability Tests*, Minneapolis: University of Minnesota Press, 1930.

[12] Marietta Apparatus Co., Marietta, Ohio.

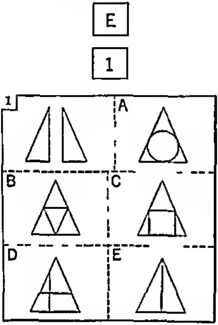
as a measure of speed and accuracy in responding to details of spatial relations and that it yields a measure of an individual's capacity to work with a variety of details in handling objects and concrete materials On the other hand, it is not adequate for measuring resourcefulness in solving problems of a mechanical nature, nor for measuring capacity to manipulate small objects with precision

*The Revised Minnesota Paper Formboard* (1948) is, as its name indicates, a test which reproduces in printed form the same type of problems as those presented by actual formboards [13] In each problem, the subject is shown two or more parts of a geometric figure, when correctly assembled, the parts will make the complete figure It is the subject's task to identify the correctly assembled figure from among five choices This test is designed, it appears, to measure one's capacity to visualize and imaginally manipulate geometric forms The data reported for it indicate that it has value for the prediction and measurement of mechanical ability and for differentiating between various grades of proficiency therein Reported research has shown this paper formboard test to have a moderately good correlation with quality of mechanical performance and a moderate to low correlation with success in mechanical drawing and descriptive geometry As a group, students in engineering and mechanical vocations obtain higher scores than do other groups of students Available evidence, however, demonstrates that this test does not have high enough predictive value to be used exclusive of other criteria and information

A number of "pencil and paper" tests are designed to evaluate mechanical aptitude by testing for specific mechanical information, selected vocabulary, and ability to perceive and deal with practical mechanical problems One of the earliest of these is the *Stenquist Mechanical Aptitude Test* (1921) [14] One part consists of problems presented by means of pictures In each instance, the subject is required to determine which of five pictures belongs with each of five others This part is, essentially, a test of the subject's knowledge about mechanical tools, objects, and devices, although there is some room for the perception of relations and for reasoning The second part of the Stenquist test consists of some material similar to that in the first section (that is, matching missing pieces with the correct mechanical

[13] Psychological Corporation New York.
[14] World Book Co, Yonkers, N Y

objects) plus questions applied to cuts of machines and machine parts One underlying assumption of a test like the Stenquist must be that knowledge about mechanical tools, objects, and devices reflects mechanical interests and that mechanical interests are indicative, in some degree, of mechanical aptitude




*First look at Problem 1 There are two parts in the upper left hand corner Now look at the five figures labelled A B C D E You are to decide which figure shows how these parts can fit together Let us first look at Figure A You will notice that Figure A does *not* look like the parts in the upper left hand would look when fitted together Neither do Figures B C or D Figure E *does* look like the parts in the upper left hand corner would look when fitted together so E is PRINTED in the square above $\boxed{1}$ at the top of the page

FIG 12 9 *Specimen Item from Revised Minnesota Paper Form Board Test* Psychological Corporation (By permission )

The *Tests of Mechanical Comprehension*[15] present mechanical problems in pictorial form In each instance, accompanying the picture is a statement of the problem depicted, with two or three answers from which to choose the correct one These tests, on three levels of difficulty, are designed to measure one s understanding of the operations of physical and mechanical principles in relatively simple situations One form is designed for use with high school students, engineering school applicants, and in general with relatively untrained and inexperienced persons A second and somewhat more difficult form is intended for use with engineering-school candidates, appli-

[15] By G K. Bennett et al New York The Psychological Corporation 1940–1951

cants for technical courses or for employment in mechanical jobs The third form was devised for use with high school girls and women Since the types of items included are intended to be appropriate to the level and experience of each group of examinees, many of the items used for the women s test come within their range of household activities, involving objects and devices used in a home rather than in a shop Other items non household in character, are also utilized

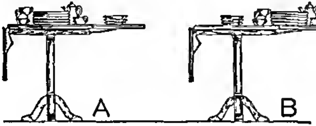



FIG 12 10 Specimen Item from Test of Mechanical Comprehension by G K Bennett and D F Fry Psychological Corporation (By permission )

Unlike some other tests of mechanical comprehension this one does not require specific knowledge, such as matching the parts of a tool or some other mechanical object, nor does it require verbal knowledge regarding tools, processes, or materials The items in the present test depict objects that are almost universal in American life, such as airplanes, ladders, stairs, wheels, gears, pulleys, see-saws, and others The hypothesis is that answers to the problems presented do not depend upon specific information or training but can, rather, be arrived at by analysis of the nonverbal materials presented The extent to which this hypothesis is satisfied varies somewhat among the 60 items of each test For example, there is no doubt that familiarity with elementary physical principles or actual experience will be an aid in answering questions involving pulleys and leverage Yet, individuals without such advantages, whose analytical ability is adequate, are often able to arrive at correct answers

*The Prognostic Test of Mechanical Abilities* (1947) is intended for guidance purposes with pupils in grades seven through twelve and for screening purposes in industry It is so devised as to provide a profile of abilities that are common to many mechanical occupations

It also provides a composite score The content of the test was determined by analyses of courses of study and by job analyses in mechanical and related occupations

The subtests are the following arithmetic computation, from simple addition through fairly simple fractions (no arithmetical-problem solving), reading simple drawings and "blueprints", identification and use of common tools, spatial relationships (the paper form board), checking measurements [16]

**Evalnation of Mechanical Aptitude Tests.** *It is evident* from the descriptions of the foregoing tests that mechanical aptitude is best regarded as a complex of several functions in the measurement of which some test authors emphasize only one or two aspects while others have made their devices more comprehensive

On the whole these tests are statistically reliable Their validity, however, may be questioned, for in evaluating them, psychologists repeatedly comment on the inadequacy of validating criteria If we regard marks in high-school shop courses, scores of occupational and educational groups (mechanic versus nonmechanic), and low correlations with tests of general intelligence as criteria, then we can say that some of the available tests in this field have a fair degree of validity for purposes of educational guidance On the whole, by comparison with tests of intelligence—such as those described in earlier chapters—available tests of mechanical aptitude are inferior in respect to definition of functions to be measured, level of standardization, and predictive value in actual performance

The Wrightstone and O'Toole test—one of the more recent—may be cited for illustrative purposes Its reliability is high, being about 90 Intercorrelations of subtest scores range from 30 to 70, with a median of 55 Correlations of subtest scores with total scores vary from 52 to 77, with a median of 70 Evidence of validity is offered in two ways in part in terms of "face validity" and in part in terms of correlations with instructors' ratings in a training course for aviation mechanics In the first instance, the test's authors state that their device measures the skills prescribed as necessary in eleven mechanical occupations In the second instance, the coefficients of contingency go from 60 to 78, with a median of 67 These coefficients are among

[16] By J W Wrightstone and C E O'Toole Published by California Test Bureau, Los Angeles

the highest in this area of investigation, and considerably higher than most others

Numerous studies have been published in which various tests of mechanical aptitude have been intercorrelated The reported coefficients are almost uniformly low (below 50) or very low A few of the reported coefficients are moderate, that is, somewhat above 50 The reasons for these relatively low coefficients—unlike those found between the sounder tests of intelligence—are to be sought in the following factors (1) Some of the tests are much more comprehensive in scope than others that are relatively restricted and homogeneous in content, hence, the former measure a greater number of functions, some of which may have little communality with the latter (2) Not all of these tests are calibrated for the same levels of difficulty, hence they do not have equal differentiating value at a given level (3) Some of the tests are much more dependent upon experience and specialized information than are others (4) Performance levels on several tests may reflect different degrees of interest and motivation in special areas

In connection with (1), above, study of the content of tests of mechanical aptitude shows that they sample, more or less, the following functions visual motor integration, spatial visualization, perceptual speed, manual dexterity, and visual insights (analysis) In addition to these, some of the tests measure specialized information, knowledge of techniques, arithmetical problem-solving ability, and technical vocabulary Some of the functions are measured by means of apparatus tests (Figures 12 4, 12 5, 12 6), others by means of performance-type materials (formboards, etc ), and still others by means of pencil and paper tests It is not surprising, therefore, that intercorrelations between these tests are low, even though they fall within the same category

On the whole, tests of mechanical aptitude show very moderate or only low correlations with actual job performance This fact does not necessarily signify that the tests themselves are defective, for many non-mechanical factors enter into the job ratings received and into actual performance on the job These factors include subjective judgments of the raters, the worker's health, motivation, and personality traits which may facilitate or impede performance

Although the foregoing factors might lower the correlation coefficients, it is highly improbable that they are solely accountable for

the relatively low relationships found It is essential, therefore, that a given test of mechanical aptitude be studied for each type of occupation where it is to be used, in order to establish its validity not only in terms of correlation coefficients but—and more significant—in terms of critical cut-off scores and expectancy tables for each of several levels of test performance (See the discussion of these techniques in Chapter 1 )

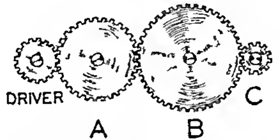



FIG 12 11 "Which gear will make the most turns in a minute?" From the Bennett Test of Mechanical Comprehension The Psychological Corporation (By permission )

Some of the tests of mechanical aptitude merit consideration when vocational or educational guidance is the problem at hand, for they are valuable as supplements to other types of information For example, the Bennett and Fry Test of Mechanical Comprehension appears to be quite useful for selection in situations where understanding of machines is necessary In any case, in guidance work it would be desirable to administer more than one test of mechanical aptitude, since intercorrelations of the several instruments are not nearly high enough to warrant their use interchangeably The particular combination of tests used in any given situation should depend upon the nature of the problem presented by the individual concerned and upon the kinds of jobs under consideration

## TESTS OF CLERICAL APTITUDE

Clerical aptitude, like mechanical is not a unitary function The tests consist of several kinds of items some of which correlate quite highly with scores of general intelligence tests, but differ from the latter in that they contain only selected materials which are of significance for clerical occupations

*The Psychological Corporation General Clerical Test* (1950) is intended to examine routine clerical aptitude, proficiency in arithmetic, and verbal facility The first is measured in terms of perception of similarities and differences between paired names and numbers, and in terms of skill in using a filing scheme The second includes the usual arithmetic processes and problems The third is tested through spelling, paragraph meaning, comprehending directions, word meaning, and language usage

1. 9 3 7 - 9 3 7 ( )
2. o b h - o b h ( )
3. Curtis & Co. - Curtis & Co. ( )
4. - ( )

57. 6 8 1 9 0 0 2 3 4 1 - 6 8 3 9 0 0 2 3 4 1 ( )
58. v r x o a e d i q f - v r x o a e d i q f ( )
59. H. W. Hieronymous - H. W. Hiernymous ( )
60. - ( )

FIG 12 12 Rate and Accuracy in Perceiving Similarities and Differences From Detroit Clerical Aptitudes Examination Public School Publishing Company (By permission )

*The Detroit Clerical Aptitudes Examination* (1944) [17] attempts to measure more comprehensively than the foregoing, although the superiority of the Detroit test has not been demonstrated The authors state that their test is intended to select pupils with capacities for commercial courses in high school The following parts are included rate and quality of handwriting, rate and accuracy in checking (perceiving similarities and differences between paired series of digits, letters, names, and small, simple geometric figures), simple arithmetic processes, motor speed and accuracy (placing X's in a page full of small circles), knowledge of simple commercial terms, visual imagery (disarranged parts of a picture), rate and accuracy of classification

[17] By H J Baker and P H Voelker, Bloomington, Ill Public School Publishing Co

(a letter number substitution test a variant of the digit symbol test) and alphabetical filing

*The Minnesota Clerical Test* (1946) is one of the most widely used in this area of aptitude testing It consists of two subtests Numbers and Names The former consists of paired numbers only some of which are identical and are to be marked by the testee for example 9632–9632 5179 5719 The Names test is designed to measure the same functions (perception of detail and perceptual speed) using identical and non identical paired names The differences within the latter pairs are minor for example Braun and Co Brown and Co These two subtests have much in common regarding psychological functions ($r = 65$) but the Names test is more inclusive and complex, for it correlates considerably higher with intelligence scales than does the Numbers test

**Evaluation of Clerical Tests** The three tests briefly described above will suffice as examples of this type for they are quite representative On the whole results obtained with clerical aptitude tests if critically interpreted may contribute to a better understanding of a pupil s capacities and to his guidance in the selection of a high-school course although the tests correlate only moderately with marks in commercial courses (from about 30 to about 50) But moderate correlations between test results and school marks are the rule rather than the exception for the size of the coefficients is affected by several factors besides inadequacies of the tests themselves namely pupils lack of interest or incentive in their courses interference of non intellectual factors such as poor health economic pressure emotional forces and extracurricular activities and the variability of teachers marks

While their *rel ability* coefficients are generally within the satisfactory range tests of clerical aptitude do not provide sufficient evidence of their *general* value for the prediction of competence and quality of performance on the job itself *Val dity* correlations generally fall between about 20 and 45 However as so often happens even when validity correlations are low the tests are useful in identifying those persons at the higher levels and those at the lower Here again cut-off scores and expectancy tables can be more revealing than correlation coefficients

Authors of these tests provide among other data norms for a variety of groups—e g high school seniors in commercial courses

office workers non-office workers, different classes of clerical workers employed and unemployed office workers—which indicate the trend in scores that would be expected for each of several groups But the range of scores within groups and the extent of overlapping of scores between groups are so considerable, and correlations are low enough so that in any individual case detailed analysis of test results must be evaluated, not in isolation, but together with other information concerning the individual under study

DIFFERENTIAL APTITUDE TESTS[18]

Rationale This battery of tests is a most ambitious and comprehensive instrument Its statistical data on standardization and analysis—reliability, validity, intercorrelations of parts norms, population samples—are exceptionally thorough

Several guiding principles were applied in the development of the parts of this battery

All eight parts of the battery were standardized on the same population Thus the norms and percentile values for each test have the same relative significance as those for all the other tests in the battery, for the ranges of age, aptitude, school grade, and non intellective personality factors were constant in the standardization process Psychological profiles therefore are more meaningful for interpretation of differences *within* an individual The published norms are based upon a population sample of 47,000 boys and girls in grades 8 through 12, from communities throughout the country Separate batteries and norms are available for boys and girls

Each test in the battery should be relatively independent and measure a relatively restricted range of aptitude

Each test in the battery should be intended for the same purpose as every other one namely, educational and vocational guidance (rather than, for example, selection for a particular job)

Each test in the battery should be useful for guidance in a number of related areas rather than in only one or two

Each test should measure *level* of aptitude—that is, power rather than speed of performance

The battery of tests should yield a profile in percentile ranks All eight percentile ranks for an individual will be comparable since they have been derived from the same population sample

[18] By G K Bennett H G Seashore A G Wesman The Psychological Corporation 1947 *Manual* Second Edition 1952

(a letter-number substitution test, a variant of the digit-symbol test), and alphabetical filing

*The Minnesota Clerical Test* (1946) is one of the most widely used in this area of aptitude testing It consists of two subtests Numbers and Names The former consists of paired numbers, only some of which are identical and are to be marked by the testee, for example 9632–9632, 5179–5719 The Names test is designed to measure the same functions (perception of detail and perceptual speed), using identical and non identical paired names The differences within the latter pairs are minor for example, Braun and Co –Brown and Co These two subtests have much in common regarding psychological functions ($r = 65$), but the Names test is more inclusive and complex, for it correlates considerably higher with intelligence scales than does the Numbers test

**Evaluation of Clerical Tests.** The three tests, briefly described above, will suffice as examples of this type, for they are quite representative On the whole, results obtained with clerical aptitude tests, if critically interpreted, may contribute to a better understanding of a pupil's capacities and to his guidance in the selection of a high-school course, although the tests correlate only moderately with marks in commercial courses (from about 30 to about 50) But moderate correlations between test results and school marks are the rule rather than the exception, for the size of the coefficients is affected by several factors besides inadequacies of the tests themselves namely, pupils' lack of interest or incentive in their courses, interference of non-intellectual factors, such as poor health, economic pressure, emotional forces, and extracurricular activities, and the variability of teachers' marks

While their *reliability* coefficients are generally within the satisfactory range, tests of clerical aptitude do not provide sufficient evidence of their *general* value for the prediction of competence and quality of performance on the job itself *Validity* correlations generally fall between about 20 and 45 However, as so often happens even when validity correlations are low, the tests are useful in identifying those persons at the higher levels and those at the lower Here again, cut-off scores and expectancy tables can be more revealing than correlation coefficients

Authors of these tests provide, among other data, norms for a variety of groups—e g, high school seniors in commercial courses,

office workers, non-office workers, different classes of clerical workers, employed and unemployed office workers—which indicate the trend in scores that would be expected for each of several groups But the range of scores within groups and the extent of overlapping of scores between groups are so considerable, and correlations are low enough, so that in any individual case detailed analysis of test results must be evaluated, not *in isolation, but together with other information* concerning the individual under study

DIFFERENTIAL APTITUDE TESTS[18]

**Rationale.** This battery of tests is a most ambitious and comprehensive instrument Its statistical data on standardization and analysis—reliability, validity, intercorrelations of parts, norms, population samples—are exceptionally thorough

Several guiding principles were applied in the development of the parts of this battery

All eight parts of the battery were standardized on the same population Thus the norms and percentile values for each test have the same relative significance as those for all the other tests in the battery, for the ranges of age, aptitude, school grade, and non-intellective personality factors were constant in the standardization process Psychological profiles, therefore, are more meaningful for interpretation of differences *within* an individual The published norms are based upon a population sample of 47,000 boys and girls in grades 8 through 12, from communities throughout the country Separate batteries and norms are available for boys and girls

Each test in the battery should be relatively independent and measure a relatively restricted range of aptitude

Each test in the battery should be intended for the same purpose as every other one namely, educational and vocational guidance (rather than, for example, selection for a particular job)

Each test in the battery should be useful for guidance in a number of related areas, rather than in only one or two

Each test should measure *level* of aptitude—that is, "power" rather than speed of performance

The battery of tests should yield a profile in percentile ranks All eight percentile ranks for an individual will be comparable since they have been derived from the same population sample

[18] By G K Bennett, H G Seashore A G Wesman The Psychological Corporation 1947 *Manual* Second Edition, 1952

**Contents.** The battery includes the following eight tests

*Verbal reasoning* verbal analogies of the familiar type are used to measure ability with more or less complex verbal concepts and relationships Since this test encompasses only a very restricted area of verbal reasoning "verbal analogies" would be a more appropriate name

*Numerical ability* numerical relationships and facility with number concepts are tested—that is, essentially computation rather than arithmetical reasoning (solving problems) Some of the items test only skill in the four fundamental processes, others require understanding of numerical concepts and relationships

*Abstract reasoning* ability to reason with nonverbal materials The series presented in each problem requires insight into an operating principle in the changing diagrams (See Fig 12 13 )

*Space relations* ability in spatial visualization is tested by presenting two-dimensional geometric figures (variously shaded) which are imaginally manipulated, each to form a three-dimensional figure Using two-dimensional drawings, the purpose is to test aptitude to visualize constructed figures, variously rotated

*Mechanical reasoning* mechanical comprehension is tested with the same types of items as those already described under *Tests of Mechanical Comprehension*

*Clerical Speed and Accuracy* speed and accuracy of responses to letter and number combinations are measured This test involves the

*Each row consists of four figures called Problem Figures and five Answer Figures The four Problem Figures make a series You are to find out which one of the Answer Figures would be the next or the fifth one in the series*


*Study the position of the black dot Note that it keeps moving around the square clockwise upper left corner, upper right corner, lower right corner lower left corner In what position will it be seen next? It will come back to the upper left corner Therefore* B *is the answer, and you would mark your Answer Sheet like this* ⟶ A B C D E

FIG 12 13 Space problem ( Abstract Reasoning ) from *Differential Aptitude Tests* The Psychological Corporation (By permission )

matching of various combinations, emphasizing perception of detail and rate of response

TEST ITEMS

| | | | | | |
|---|---|---|---|---|---|
| V | AB | AC | AD | AE | AF |
| W. | aA | aB | BA | Ba | Bb |
| X | A7 | 7A | B7 | 7B | AB |
| Y | Aa | Ba | bA | BA | bB |
| Z. | 3A | 3B | 33 | B3 | BB |

SAMPLE OF ANSWER SHEET

| | | | | | |
|---|---|---|---|---|---|
| V | AC | AE | AF | AB | AD |
| W | BA | Ba | Bb | aA | aB |
| X | 7B | B7 | AB | 7A | A7 |
| Y | Aa | bA | bB | Ba | BA |
| Z | BB | 3B | B3 | 3A | 33 |

FIG 12 14 Number and letter test items from differential aptitude tests In each row of the *Sample Answer Sheet*, the examinee marks the combination that matches the underlined combination in the corresponding *left* hand row The Psychological Corporation (By permission )

*Language usage* Part I is simply a spelling test Some words are correctly spelled, others are incorrectly spelled The subject indicates for each word whether it is right or wrong To call this test "language usage" is to stretch the meaning of that term to unreasonable lengths

*Language usage* Part II is made up of sentences in which the examinee is required to distinguish faulty from correct grammar, punctuation, and word usage

**Evaluation.** As already stated, these are among the few most thoroughly standardized and reported aptitude tests available They measure several psychological functions that have been shown to be significant in educational and vocational guidance In this connection, the intercorrelations of the eight parts show that each has a separate and different—though not entirely unique—contribution to make, for the coefficients range (for boys) from 06 (Mechanical Speed with Clerical Speed and Accuracy) to 62 (Sentences with Verbal Reasoning) The range of coefficients for girls is approximately the same ( 12– 67)

Extensive correlational studies with other standardized tests, of a variety of types, show that the several parts of the DAT do not measure entirely the same functions as do these others, although some of the parts (Verbal Reasoning and Numerical Ability) are often highly correlated with tests of general intelligence The DAT, therefore, has a contribution to make in a comprehensive study of an individual s aptitudes

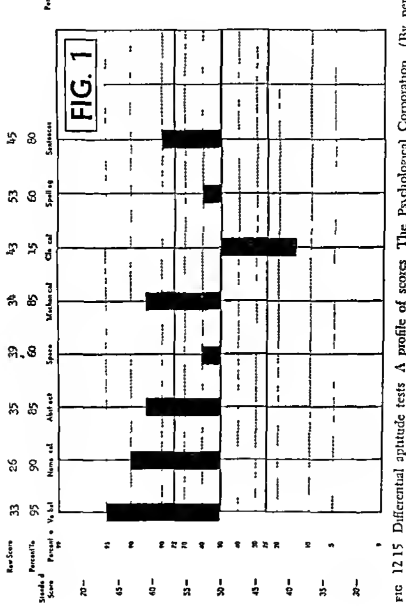



FIG 12 15 Differential aptitude tests A profile of scores The Psychological Corporation (By permission)

Extensive data are also reported showing validity coefficients between DAT scores and grades in a variety of school subjects These indexes are of approximately the same magnitude as those found with other sound tests The correlations on the parts of the DAT vary with the several school subjects used as validity criteria, so that, as is to be expected some parts have greater predictive value in a given area of studies than do others

In addition to the foregoing, norms of performance are provided for a number of different educational and occupational groups

The tests are adequately reliable for the most part, the mean coefficients for the several grade levels varying from 85 to 93 for boys, and 71 to 92 for girls The test of Mechanical Reasoning (reliability = 71) is the only part that is seriously below accepted degrees of reliability

The scores of this battery are reported in profile (graph) form, as shown in Figure 12 15 The interpretation of such a profile and the significance to be attached to score differences as shown on the graph depend upon adequate comprehension of the statistical evidence and the psychological rationale presented in the Manual and both of these will make more than the usual demands upon the non specialist's technical sophistication

As is the case with all tests of aptitudes the ultimate value of the DAT will depend upon follow up studies showing their predictive efficiency in learning and job performance For this purpose, the authors and users of this battery of tests will have to determine the differential and predictive significance of profile "patterns" and multiple correlations (between two or more parts of the tests, on the one hand and the criterion on the other), thus going beyond the simple correlation coefficients and expectancy tables, important though these are

## APTITUDE CLASSIFICATION TESTS [19]

This battery of tests has been devised for purposes of vocational guidance Its recommended occupational range (thirty vocations) covers a wide area at several levels including such different

[19] By J C Flanagan Science Research Associates 1953 This test is described at this point because it is intended for use in vocational guidance covering a range of occupations from clerical and mechanical through some professions. It may therefore serve as a bridge to the following chapter

occupations as office clerk, mechanic, humanities professor, nurse, physicist, and writer

The battery comprises fourteen tests, each of which is briefly described below

inspection ability to spot flaws
coding speed and accuracy in coding typical office information
memory recall of codes learned in coding test
precision speed and accuracy in making small circular finger movements
assembly ability to visualize the appearance of an object from a number of separate parts
scales speed and accuracy in reading scales, graphs, and charts
coordination ability to coordinate hand and arm movements
judgment and comprehension ability to read, understand, reason, and use good judgment
arithmetic the four fundamental processes
patterns ability to reproduce simple pattern outlines
components ability to identify important component parts in line drawings and blueprint sketches
tables reading two types of tables—one using numbers the other using words and letters
mechanics understanding mechanical principles
expression communicating ideas in writing and talking

This device is of the differential aptitude type Its author suggests various combinations of the fourteen tests as having prognostic value in the selection of a vocation For example, for the prospective occupation of nurse, the three recommended tests are (1) memory, (2) scales, (3) judgment and comprehension, for the prospective humanities professor the recommended tests are (1) memory, (2) judgment and comprehension, (3) expression So far as this battery is able to differentiate between aptitudes, then, the distinction between these two widely diverse professions appears to depend upon the tests of expression and of scales It is hardly probable, however, that the differences in aptitudes, in this instance, are as simple and limited as these sets of tests would indicate Nor is it probable that valid selections for either occupation alone can be made on the basis of these few and simple tests The author of the test battery recognizes this fact, for he states that the test combinations are only tentatively proposed, and he recognizes that tests for the measurement of additional "job elements" may be necessary Similar comparisons can be made

between many pairs of the thirty occupations for which test combinations are recommended

So far as statistical analysis of standardization data is concerned, the results are fairly satisfactory Intercorrelations between individual tests are generally low seven of ninety one coefficients are 50 or higher, the median is 29 Standard errors of measurement are moderate Reliability coefficients for individual tests are fair ( 26 to 86, Form A, 55 to 85, Form B) However, reliability coefficients for *combined* scores of all tests recommended for a particular occupation are much higher for nine occupations reported namely, from 83 to 93 These reliability data make it clear that the "job element approach," upon which this battery is based, yields much less stable results than does a more nearly "wholistic" or "global" approach

So far as predictive (on-the-job) validity is concerned, the value of these tests remains to be demonstrated The standardization population consisted of high-school seniors Only a few of these have been followed-up and rated in their occupations The correlations between vocational test scores and occupational progress after four years (rate of salary increase), in seven occupations, ranged from 29 to 64 Correlations of test scores (obtained in senior year of high school) with college grades were from 24 to 36

It is clear that in their present state, the Aptitude Classification Tests must be regarded as tentative efforts in the direction of vocational guidance by means of tests based on the "job element approach"

# 13.

## APTITUDE TESTS: FINE ARTS AND PROFESSIONS

### *TESTS OF MUSICAL APTITUDE*

The Seashore Tests. The most widely known are the *Seashore Measures of Musical Talent* (1919–1939)[1] The six parts, recorded on phonograph records, are devised to measure (1) pitch discrimination (graded from readily discernible differences to very fine ones), (2) intensity or loudness discrimination (pairs of clicks, with gradation differences), (3) time discrimination (differing time intervals between clicks), (4) discrimination of timbre, or tone quality (pairs of tones differing in quality), (5) rhythm discrimination (*pairs of rhythmic patterns of increasing complexity*, to be discerned as the same or different), and (6) tonal memory (paired tonal patterns requiring perception of differences between members of the pair) There are two sets of records one for unselected groups and the other for musicians and students of music Total scores for the six parts do not represent an individual s test performance, instead profiles of the six auditory capacities should be used and evaluations should be based upon them

Seashore approached the development of his measures from a point of view different from that of authors of most other types of aptitude tests Instead of making a "job analysis," which would attempt to discriminate levels of musical aptitude on an empirical basis, Seashore made what he regarded as a theoretical analysis of musical talent

[1] RCA Manufacturing Co Camden, New Jersey

into its sensory components Some of these components, he held, can be measured objectively, whereas others cannot The six capacities mentioned above are among the measurable ones, but as Seashore himself insisted, they do not offer a complete index or profile of all components of musical aptitude They are measures only of auditory perception required in music, analyzed into six components

For the purpose of validation, results of the Seashore measures have been compared with teachers' ratings of musical ability, with musical achievement, and with quality of work in schools of music. The obtained correlations have not been such as to warrant the conclusion that these tests of auditory perception are sufficiently valid in predicting various levels of musical talent Seashore himself, however, has objected to attempts at an over-all validation of his measures While he has not done so himself, he maintains that each of the six tests should be separately validated against different kinds of specialized musical activity For example, the test of pitch discrimination should be validated especially for players of string instruments At any rate, this much may be said: the Seashore measures reveal those persons whose auditory perceptions are so deficient that they could not successfully participate in the formal study or performance of music.[2]

**The Wing Tests.** *The Wing Standardized Tests of Musical Intelligence* (1948),[3] also presented by means of a series of recordings, is intended to meet at least two objections that frequently have been directed against the Seashore tests namely, that the latter are "atomistic" and that they are not based upon functions being trained and considered most important by teachers of music The Wing tests measure seven aspects of musical perceptiveness, yielding a score for

[2] See J C Saetveit, D Lewis, and C E Seashore, *Revision of the Seashore Measures of Musical Talent* Iowa City University of Iowa 1940, H M Stanton, *Measurement of Musical Talent* University of Iowa 1935, C E Seashore, *Psychology of Music* New York McGraw Hill Book Co, 1938, J L. Mursell, *The Psychology of Music* New York W W Norton, 1937, Paul R Farnsworth, "An Historical Critical, and Experimental Study of the Seashore Kwalwasser Test Battery," *Genetic Psychology Monographs*, Vol 9, 1931, pp 291–393

[3] By H D Wing and C Wing Sheffield City Training College, Sheffield, England Cf O K Buros, *Fourth Mental Measurement Yearbook*, pp 344 ff The use of the term 'musical intelligence' is unfortunate, since the word "intelligence" has been given another psychological connotation

as are sampled by the Wing tests [6] This method of testing is indicated especially in view of the fact that teachers of music and others in that field maintain that the separate sensory functions (as measured in the Seashore) do not operate in music as they do in the controlled and isolated test situation

## TESTS OF APTITUDE IN THE GRAPHIC ARTS

**The Knauber Test.** *The Knauber Art Ability Test* (1932–35),[7] for use in grades 7 to 16, requires the following drawing a design from memory, drawing from memory figures within space limitations, drawing a stereotyped character such as Santa Claus, arranging within a given space a specified composition, creating and completing designs from supplied elements, spotting errors in drawn compositions, such as incorrect perspective, misplaced details, incorrectly proportioned details, incongruous or inconsistent elements, production of compositions intended to show creative imagination, ingenuity, ability to represent a concept symbolically, or to plan and execute a universal idea

Attempted measurements of these complex capacities represent a large undertaking for a single test Performance on some parts, it appears, is dependent upon mastery of stereotypes and traditional problems and tasks presented in art instruction This being the case, the Knauber test would be useful primarily for evaluating school progress in art, quality of observation, and, to some extent creative imagination That is, the test would be in large part a measure of learning rather than a measure whose primary usefulness is with individuals who have not had the benefits of some formal instruction

**The Meier Test.** *The Meier Art Judgment Test* (1940),[8] for grades 7 to 12, consists of 100 pairs of uncolored pictures, one member of each pair being a reproduction of a masterpiece The pictures cover a wide variety of subjects landscape, still life, woodcuts oriental drawing murals, and others The second member of the pair is altered

---

[6] See J McLeish "The Validation of Seashore s Measures of Musical Talent by Factorial Methods" *British Journal of Psychology Statistical Section* Vol 3 1950 pp 129–140 Also H D Wing "A Factorial Study of Musical Tests" *British Journal of Psychology* Vol 31, 1941 pp 341–355

[7] Psychological Corporation New York

[8] Bureau of Educational Research and Service State University of Iowa, Iowa City

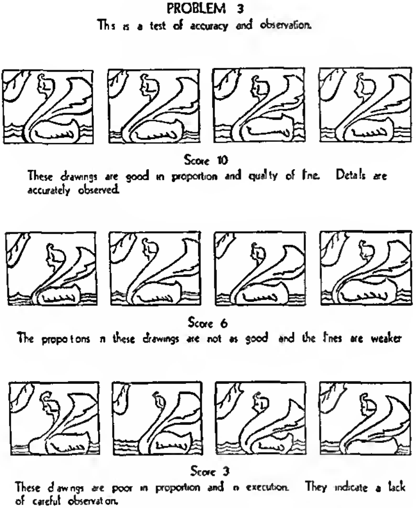



FIG 13 1 *The subject is shown a design in the examination booklet which he is asked to copy This item is a test of accuracy and quality of reproduction From the Knauber Art Ability Test (By permission )*

from the original in some respect and in such a manner as to make it inferior to the original The nature of the alteration in each pair is indicated (perspective, use of curves, arrangement, etc ), the subject is, of course, not told which one is the original and which is the altered copy It is his task to indicate his preference as between the two copies of each pair The author of this test has selected not artistic execu-

tion but, rather, aesthetic judgment as the principal factor in art aptitude and as the only factor to be tested, the contention being that judgment is the "key-capacity," the most trustworthy and significant index to talent in art and to success in a career in art The soundness of this assertion, however, remains to be validated If it were correct that artistic perception and judgment constitute the "key-capacity,"

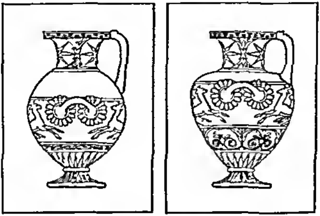


FIG 13 2 *One of these pictures represents an artistic work of established merit The other is an adaptation of that work, and aesthetically inferior In this pair, the subject is required to select the original and aesthetically superior work on the basis of the shapes of the bowls From the Meier Art Judgment Test* Bureau of Educational Research and Service State University of Iowa (By permission )

rather than actual execution and creativity, we should then expect art critics and writers on art to be the most talented in actual production, but such is not the case

**The Graves Test** The *Graves Design Judgment Test* (1948) [9] is devised to measure appreciation and production of art The author enumerates and describes eight principles that he regards as basic namely, unity, dominance, variety, balance, continuity, symmetry, proportion, and rhythm The purpose of the test is to evaluate the extent to which an individual is able to apprehend and respond to these principles

[9] Psychological Corporation, New York.

versal criteria for the selection and scoring of test items and for the validation of test scores, once the items have been selected

Closely related to the foregoing problem is the inevitable result that different teachers and critics of art, employing varying criteria in their evaluations, will rate art productions differently The result is that there are relatively few useful data on the predictive efficiency (validity) of tests in the fine arts with regard to level and quality of performance in educational and vocational activities Where coefficients are reported, they are only fair to low, when correlated with marks in art courses or with teachers' estimates of students' abilities At the same time, however, available tests are useful in discovering individuals of unusual capacity and, in general, in evaluating appreciation of some of the graphic arts These tests, also, quite consistently differentiate between art students and non-art students, the differences between their mean scores being significant

For non art students, the tests can be useful in evaluating capacities that are essential in aesthetic appreciation, as a basis for non-vocational education in the fine arts This is true even though the capacities measured by the tests, though essential, are not sufficient in themselves to indicate subsequent quality of creative or even only reproductive performance

Finally, the available tests of aesthetic judgment assume, apparently, that this capacity is generalized and transferable from evaluations of pictures and of relatively pure designs to architecture, furniture, clothing, sculpture, and industrial products This assumption may be warranted Actual demonstration of its validity would be of significance in school and college courses that offer instruction designed to cultivate aesthetic judgment Definitive research has yet to be done on this psychological and educational problem

### TESTS OF APTITUDE IN MEDICINE[10]

**General Characteristics.** Several tests of medical aptitude have been constructed The first one has appeared in 25 revised forms over a period of about twenty years, under the sponsorship of the Committee on Aptitude Tests for Medical Students of the Association of American Medical Colleges, and under direction of F A Moss

[10] See I L. Kandel, *Professional Aptitude Tests in Medicine Law, and Engineering* New York. Teachers College, Columbia University, 1940

Recent editions have included all or most of the following subtests visual memory, memory for content, scientific vocabulary, understanding of printed material, scientific definitions, and logical reasoning

These tests have been based upon an analysis of the qualifications necessary for the successful study of medicine, which are given as the following First, it is essential to have sufficient mental alertness to learn quickly and to organize the material learned so that it can be retained and utilized in later work A sampling of medical materials was used in the test to examine this capacity Second, past scholastic performance may be expected to indicate future learning Inasmuch as all premedical students have had elementary courses in chemistry, physics, biology, and English sections of the test are devoted to questions in these subjects to determine the extent of the candidates' learning in them Third, the capacity to make correct interpretations and deductions from given data was considered essential Hence, the test included a passage of difficult reading, dealing with materials found in medical studies The testee is required to make certain interpretations and deductions based upon the passage, to which he may refer at any time Fourth, since medical students and practicing physicians are expected to draw conclusions and make diagnoses from given facts, a subtest was devised to evaluate logical reasoning " This consists of a set of premises and conclusions drawn from them, the student's task being to determine whether or not the conclusions are warranted

The reader has surely noted that the mental capacities to be measured by means of the foregoing types of items are by no means peculiar to the study of medicine They are, indeed, capacities which are required in all fields of study and in all professions The tests of this medical committee, therefore, though they are called aptitude tests, are actually tests of general ability, *utilizing in part a special content* which is included in or closely associated with the materials of study in medical schools In other words, the *form* of mental activity being tested is the same as in any other professional field, but the *content* is in part specialized It is in this sense that this test and others like it are aptitude tests, as that term has been defined It is important to note this fact, for otherwise one might get the impression that a specialized aptitude, independent of general ability, is required for the study of medicine

**Contents.** At present, the *Medical College Admissions Test* (1946–1951) is being further developed and administered by the Educational Testing Service[11] for the Association of American Medical Colleges The purpose of this aptitude test was stated to be to provide highly dependable measures of the advanced student s *general ability* and of his *achievement in a special field of study* The tests are predicated upon the principle that a significant aspect of potentiality for a specialized field of study at the graduate and pre professional level may be measured by testing the candidate s general scholastic ability and his achievement in a special field which is prerequisite to advanced study in the same or a closely related field

In addition, a test of understanding modern society has been included in the battery This part includes current history, economics, political science and sociology, the purpose being to evaluate alertness to social issues rather than materials retained from college courses in these fields of study

To measure *general ability* the battery has two major divisions verbal and quantitative Verbal ability is measured by tests of vocabulary (word opposites), sentence completion, and word analogies Specimens of these follow

*Opposites*

Harmony (a) accord (b) affinity (c) oppression (d) conflict (e) desecration

*Sentence completion*

The manufacturing of small machinery is profitable for the nation lacking mineral resources inasmuch as __________ can be exploited in place of __________

(a) quantity-quality (b) power-efficiency (c) skills-quality (d) alloys-ores (e) skills materials

In addition to the foregoing, the test on current society and the test on materials from college courses in basic sciences may properly be regarded as contributing to an evaluation of the candidate s general verbal ability although these two are not intended primarily for this purpose The reported correlation of 77 between verbal test scores and scores on understanding modern society supports this view, as do the other intercorrelations between the nonquantitative parts (all above 65)

[11] Princeton, New Jersey This test was formerly called the Professional Aptitude Test.

Ability in quantitative thinking is measured with tests of arithmetical reasoning applied and abstract

The test of achievement in a specialized field of study is, naturally, a science test covering a wide sampling of concepts and problems from basic college courses in biology, chemistry, and physics These items are intended to evaluate the candidate s grasp of fundamental principles of science "

The science test includes completions, classifications, analogies, quantitative comparisons, and paragraph comprehension Taken as a whole the science test is concerned with basic scientific concepts, and with applications and problems in the several fields The problems are of increasing complexity requiring comprehension, interpretation, inference, and analysis of data—in general, the use of knowledge in dealing with multi phase problems Sample items and descriptions follow

COMPLETIONS

*Sample Directions* Each of the following incomplete statements is followed by five suggested completions Select the one completion which is best in each case and indicate your selection in the appropriate space on the answer sheet

16 A sodium atom and a sodium ion
- (A) contain the same number of electrons
- (B) contain the same number of protons
- (C) have the same chemical properties
- (D) have the same physical properties
- (E) have different atomic numbers

CLASSIFICATIONS

*Sample Directions* Each of the numbered words or phrases below is associated with one both or neither of the headings listed as (A) and (B) above it On the appropriate line of the answer sheet blacken the space under

- A if the numbered word or phrase is associated with (A) *only*
- B if the numbered word or phrase is associated with (B) *only*
- C if the numbered word or phrase is associated with *both* (A) *and* (B),
- D if the numbered word or phrase is associated with *neither* (A) *nor* (B)

*Sample Questions*

(A) Thyroid gland
(B) Pituitary gland

(C) Both
(D) Neither

17 Giantism
18 Low blood calcium
19 Cretinism
20 Short stature

ANALOGIES

*Sample Directions* Each of the following questions consists of an in complete analogy with five suggested completions Select the one word or phrase which best completes the analogy and indicate your selection in the appropriate space on the answer sheet

*Sample Questions*

21 ohm resistance watt ________
(A) electricity (B) work (C) power (D) current
(E) potential

22 atom molecule *element* ________
(A) electron (B) mixture (C) isomer (D) isotope
(E) compound

23 yolk egg ________ bean seed
(A) hypocotyl (B) epicotyl (C) cotyledon (D) testa
(E) endosperm

*QUANTITATIVE COMPARISONS*

*Sample Directions* The following paired statements describe two entities which are to be compared in a quantitative sense On the appropriate line of the answer sheet blacken the space under

A if A is *greater than* B
B if B is *greater than* A,
C if the two are *equal or very nearly equal*

*Sample Questions*

24 (A) The total resistance of two given resistances in series
(B) The total resistance of the same two resistances in parallel

25 (A) The volume occupied by one gram molecular weight of helium at standard conditions
(B) The volume occupied by one gram molecular weight of oxygen at standard conditions

26 (A) The concentration of oxygen in the right auricle of a mammalian heart
(B) The concentration of oxygen in the left auricle of a mammalian heart

PARAGRAPH COMPREHENSION

*Sample Directions* In this part of the test there are several passages, each followed by a series of statements. Read the passage and then

classify each of the statements under one of the following cate gories

(A) The statement is warranted by information given in the passage
(B) The statement is true but not warranted by the passage
(C) The statement is contradicted by the passage
(D) The statement is contradicted by established evidence but not by the passage
[Scientific passages then are given each followed by several statements to be marked as directed above ]

In a second part of paragraph comprehension the passages are followed by five questions from which the best answer is to be se lected based upon material in the paragraph

*Sample Passage*

The only carbohydrate which the human body can absorb and oxid ze is the simple sugar glucose. Therefore all carbohydrates which are consumed must be changed to glucose by the body before they can be used *There are specific enzymes in the mouth* the stom ach and the small intestine which break down complex carbohy drates All the monosaccharides are changed to glucose by enzymes secreted by the intestinal glands and the glucose is absorbed by the capillaries of the villi

The following s mple test is used to determine the presence of the monosaccharides If Benedict s solution is added to a solution con taining glucose or one of the other monosaccharides and the resulting mixture is heated a brick red precipitate will be formed This test was carried out on several substances and the information in the following table was obtained P ind cates that the prec pitate was formed and N indicates that no reaction was observed

| *Material Tested* | *Observation* |
|---|---|
| Crushed grapes in water | P |
| Cane sugar in water | N |
| Corn syrup | P |
| Molasses | N |

*Sample Questions on Passage*

31 From the results of the test made upon crushed grapes in water one may say that grapes contain
(A) glucose
(B) sucrose
(C) a monosaccharide
(D) no sucrose
(E) no glucose

32 The carbohydrate content of which one of the following foods probably undergoes the LEAST change during the digestive process in the human body?
 (A) Cane sugar
 (B) Corn syrup
 (C) Molasses
 (D) Bread
 (E) Potato

**Evaluation of Medical Aptitude Tests.** The purpose of these tests is to provide an improved basis for predicting quality of performance in medical *studies*, not in medical practice In some instances, the tests have shown better results than have premedical course marks, in other instances the reverse has been true But in all instances, Moss found that the best criterion is a combination of test results and premedical course marks For example, he reports in one study the following correlations premedical course marks with medical school averages, 0 67, test scores with medical school averages, 0 64, medical school averages with a combination of test scores and premedical marks, 0 81 (multiple correlation) These are very satisfactory coefficients Other correlational studies yielded coefficients that were higher in some instances but much lower in others

The variation in correlation coefficients between test scores and medical school grades, found in different studies, is not attributable solely, or perhaps even principally, to defects in the medical aptitude tests The differences among coefficients must also reflect serious differences in medical school grading standards, inequalities of undergraduate preparation (which to some extent can be compensated for in medical school courses), and personality traits which tend to produce inconsistencies between promise and performance

Expectancy statistics have also provided useful findings regarding test scores For example, one study reports that only one percent of the highest decile students failed in medical school, whereas eighteen percent of the lowest-decile students failed These findings provide an argument for admitting all top-decile students but *not* for refusing admission to those in the bottom decile group Another study reported that the lowest decile group contributed 25 percent of the failures—that is, two and a half times its quota, in proportion to the total group of students

Reliabilities of the *Medical College Admissions Test* fall well

he acceptable range, namely, from 89 to 94 for the several 'hese reliability data again show that it is not as difficult to i reliability as it is to demonstrate validity

e these aptitude tests are not intended to predict effectiveness cal practice which, like other professions, is dependent upon lex of factors the tests appear also to have some value in fore- medical students levels of success in internships When a if interns were rated on a five point scale by their hospital staffs, ılts showed that the tests have some selective value, especially ifying students who prove to be the most satisfactory interns [1]

i study in professional schools, including medicine, now within ources of very large numbers of students, continued research , to the development of increasingly effective testing instruments ntial It would be desirable to have thorough studies made of ictors as effects of coaching and cramming upon the medical le test scores, relationships between test scores and ratings on t inventories, relationship of test scores to drop-outs in the of expectancy tables (that is, value of the tests in predicting al in the professional school, not in predicting grades alone), itions between test scores and interview ratings, role of person- raits in medical school performance and survival (that is, degree otional stability, degree and type of motivation, introversion- ersion, dominance submission, kinds and strengths of values, Admittedly, research on these and other personality traits would ficult and long, but they might, nonetheless, prove significant

## OF APTITUDE IN LAW

**General Characteristics** Tests in this field are aptitude tests same sense as are those in the field of medicine Psychologists rned with the problem of testing aptitude for legal study agree the following abilities are most important reading rapidly and rehending relatively difficult material, rapid memorizing and ac- e recall, reasoning by analogy, discriminating between the rele- and the irrelevant in a mass of facts, reasoning inductively and ctively, facility in using and acquiring a vocabulary Legal apti- tests thus far constructed attempt to measure most or all of these ies in some degree

---

Kandel *op cit* p 20